Spring 2013

# Developing a measure of student resume quality in student affairs assessment: An application of generalizability theory

Oksana Naumenko
*James Madison University*

# Developing a Measure of Student Résumé Quality in Student Affairs Assessment: An Application of Generalizability Theory

Oksana Naumenko

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

May 2013

Acknowledgements

I am beholden to several individuals who have contributed tremendously to my academic and personal development throughout my graduate work. Most importantly, I would like to thank my advisor and mentor, Dr. Keston Fulcher, who has inspired me to explore my current area of study. His patient guidance had kept me on the path to this project's completion and his irreplaceable insight has embellished the pages of this manuscript. Moreover, his support and kindness had given me the motivation to challenge myself and reach out to new horizons.

Further, I would like to thank my other committee members, whose expertise and enduring enthusiasm for student development had played a key role in my learning. Dr. Donna Sundre and Dr. Christine DeMars had contributed to the development of my enthusiasm for ensuring fairness and commitment to best practices in higher education inquiries and were instrumental in my understanding of key concepts within my work. I could not have accomplished this major task without these two dynamic and inspiring individuals.

I must also thank Dr. Kurt Schick, who has patiently helped to develop my writing ability. Along these lines, I must also acknowledge all the staff at the James Madison University Career & Academic Planning Office for becoming my community of support and for allowing this study to come to fruition. I also would like to thank my friends Jessica Hayden, Christina Ferrari and Sarah Gottlieb for being an incredible support system. Lastly and importantly, I must thank my parents, Marina and Oleg, who have sacrificed a tremendous amount to ensure the plentitude of opportunities for their only child.

Table of Contents

List of Tables

List of Figures

## Abstract

This thesis investigated the assessment of student résumés by a career services office. Specifically, the dependability of assessment scores was examined prior to making inferences regarding the value added by a career office's résumé appointment program. Systematic errors in performance assessment ratings of student résumés were examined to determine the overall dependability of the assessment scores and the precision with which raters score student performance. The absolute dependability of scores was excellent when rubric elements were fixed. Recommendations regarding training and measurement tool improvement were provided given information regarding rater precision around rubric element scores. Such evidence adds to the assessment scores' validity argument and, specifically, to the validity of inferences regarding the generalizability of student performance assessment scores (Kane, Crook & Cohen, 1999). The value-added analyses revealed medium to large standardized effect sizes across most résumé elements. Measurement information revealed that the area associated with no improvement – objective statement - was affected by inconsistent scoring rules.

CHAPTER ONE

## Introduction

Due to increasing accountability mandates (Ewell, 2002) and the growing influence that governing boards and legislatures exercise over institutional operations (Keeling & Dungy, 2004), higher education is increasingly obliged to change students' lives. As a result, many student affairs practitioners find themselves tasked with providing student outcome evidence. In turn, assessment plays an important role in helping practitioners implement strategic planning, and organizational effectiveness (Schuh, 2009). However, measurement issues often present challenges to practitioners in interpretation of assessment results. This thesis examines inferences made from performance assessment results in a particular type of student affairs context – a career planning office. Specifically, résumé writing assessment is addressed.

Student affairs assessment practice has burgeoned since the early 1990's (Schuh, 2009). Workshops on the subject shifted emphasis from merely practicing assessment to doing so using correct methodology (Schuh, 2009). Concurrently, scholarship on student affairs assessment has introduced bridges between effective assessment and sound quantitative methodology (e.g., Bresciani, 2009; Schuh, 2009). However, Student Affairs-related texts may include only limited information on research and quantitative methodology. Providing applied examples of methodologically rigorous, but accessible assessment studies can widen the scope of the student affairs practitioner's "toolbox." Readable accounts of using methodological and statistical techniques in assessment contexts are useful because it allows the practitioners to answer more sophisticated questions without turning to formal training. In turn, a wider scope of inferences about

student development can be introduced. More importantly, understanding and using the connections between research design, statistical analyses, and inferences we make from assessment results will inevitably lead to accurate decisions regarding student success. This thesis serves in part as an applied example of building an argument and evidence base for inferences about students' cognitive outcomes.

Although much of the cognitive measurement literature focuses on validating scores on objective measures, a growing recognition of objective test limitations has resulted in an emerging interest in performance assessment and its tools (e.g., Frederiksen & Collins, 1989; Linacre & Wright, 2002; Myford & Wolfe, 2003, 2004, 2009; Wiggins, 1989, 1993). Objective measures are instruments that exclude subjective judgment from the test scoring process. As such, objective tests present many practical advantages in testing, such as feasibility to test many examinees simultaneously, relatively short testing periods, and simple scoring (Kane, Crooks & Cohen, 1999). However, objective tests serve as proxies to educational outcomes, and more direct assessment of learning is more appropriate in many situations. In fact, the direct assessment of some of the most socially and economically valuable educational outcomes (e.g., oral and written communication) require observations of complex performances (Black, 1998; Kane, Crooks & Cohen, 1999; Shepard, 2000). Such direct assessments are commended for high fidelity between student ability and performance (Black, 1998; Fitzpatrick & Morrison, 1971). However, with benefits of more directly observing student capabilities come problems of implementation and validation (Mehrens, 1992; Messick, 1994). It follows that higher education practitioners should be aware of advantages, disadvantages, and methodologies

associated with performance assessment. This thesis addresses performance assessment methodology in a university career planning office.

Higher education institutions stress the importance of graduating individuals that are equipped for the workforce. By actively participating in educational experiences within general education and major courses, students are thought to acquire competency in cognitive skills necessary for employment (Shavelson, 2009; Suskie, 2006). Campus career service offices work with students to highlight such skills, thereby increasing employment chances in an unpredictable economy (Engelland, Workman & Singh, 2000). While many students and career guidance practitioners value these marketing-to-employer interventions, policy-makers require evidence to justify program support (Engelland et al., 2000; Maguire, 2003). Unfortunately, there is insufficient empirical evidence to conclude that career services interventions are effective (Pascarella & Terenzini, 2005). A likely origin of this problem is a deficiency in reliable evaluative measures. The current study gathers validity evidence for one in-house evaluative tool: The Résumé Ruler. As the name implies, it is designed to evaluate résumé quality. Assembling construct validity evidence for the instrument leads to accurate interpretation of student scores and thus more accurate interpretations of the effectiveness of career programs in relation to résumé writing outcomes. In coming up with the most appropriate set of evidence, one must be familiar with the context within which the instrument is used. The context within which construct validity evidence is assembled in turn guides the outline of a program's assessment plan. The context for career service assessment is presented next.

**Importance of On-Campus Career Services**

Student Affairs (SA) services represent a diverse out-of-classroom set of student learning and development opportunities**.** SA functions include housing and dining, physical and mental health care, recreation, cultural activities, sports, testing, orientation, career assistance, job placement, financial assistance, and disability services (UNESCO, 1998). Given the unpredictability of the current economic trends, arguably one of the most important SA functions is career development programming (Dey & Real, 2009; Engelland, et al., 2000).

Career services offices have been characterized as networking centers that supply comprehensive career-related services (Dey & Real, 2009) and focus on post-college employment. Comprehensive services include counseling and advising, career fairs, on-campus recruitment, and internship search (Tillotson & Osborn, 2012). Typical service interventions include academic advising, career planning courses, résumé development, and mock interviews. Often, campus career offices design college-career-planning courses that increase career decidedness in science and engineering students (Lent, Larkin & Hasegawa, 1986), undecided first-year students (Carver & Smart, 1985), and upperclassmen (Thomas & McDaniel, 2004). Career services also instruct and assess students in résumé development (Laker & Laker, 2007; Tillotson & Osborn, 2012). The multifaceted nature of career services yields many diverse learning outcomes. This thesis focuses on learning outcomes associated with a résumé-review program.

Career service utility on college campuses may be especially significant in the current economic context. In 2010, the National Center for Education Statistics (NCES) estimated that about 2.4 million graduates applied for jobs (Petrecca, 2010), yet the Congressional Budget Office projected the unemployment rate for that demographic is

likely around 10 percent during the same time period. Moreover, the 2.4 million graduates were also competing with the 15 million American in search of work: unemployed graduates from previous years, laid-off workers, and struggling retirees (Petrecca, 2010; Rampell & Hernandez, 2010; Simon, 2010). Given the recent economic struggles (Bureau of Labor Statistics, 2012) and steadily rising post-secondary education costs, one of the main goals of career service offices is effective career development instruction. In order to demonstrate and improve career development instruction, career services can engage in systematic assessment processes (e.g., Shutt, Garrett, Lynch & Dean, 2012).

**Assessing Career Programs**

Assessment of career services is important for several reasons. First, college career advisors have a diverse set of responsibilities, and their time should be spent efficiently to accommodate a large number of students in need of advice. Program effectiveness evidence helps determine which programs should be further developed and which should be abandoned. Second, positive program outcomes can increase student motivation to participate. For instance, student participation is likely to be encouraged by a career office reputation of high internship placement rates. The university is thus able to provide services for a larger number of students, creating a successful student graduate body. Finally, the most important reason for assessing career services is ensuring graduate competitiveness in our current economy. Well-executed career-related assessment is guided by previous results, allowing programs to allocate time and other resources to programming that actually improves student employability. Engelland, Workman, and Singh (2000) cite one additional reason to assess career services using

value-added approaches: "…institutions that address career development performance in their value-added assessment programs will be well ahead in the eyes of institutional stakeholders." (p. 234). Nevertheless, although it is reassuring that career development assessment is valued, assessment's mere presence does not guarantee improvement. When working towards the successful future of college graduates, it is important that the measurement of career service program learning outcomes is sound.

**Résumé Writing Improvement Programs**

Given that résumés link one's college experiences to potential jobs, they are a logical focus point – both in terms of programming and assessment - for career services. The résumé is an organized, professional profile that highlights an applicant's strengths, accomplishments, interests, skills, and work-related experiences that should be viewed as a powerful, self-marketing tool (Shakoor, 2001). Given that résumé submission is a critical part of the job procurement process, one could argue that developing a well-designed and well-constructed résumé contributes directly to employment (Ross & Young, 2005; Toporek & Flamer, 2009). A résumé is a brief summary of a person's education, job experience, and professional accomplishments (Hoheb, 2002). Aside from this, résumés also contain a person's contact information, qualifications, association memberships, and special skills, awards, and honors (Crosby, 2009). Many of these information pieces can be presented in ways that either guarantee the job or send one to the bottom of the applicant pile.

Perhaps because of its artful nature, many college students consider résumé writing an intimidating task (Ross & Young, 2005). Fortunately, the skills necessary to develop a résumé can be honed through co-curricular instruction by university career

office staff. In campus career development offices, résumé programming typically occurs in the form of workshops (Tillotson & Osborn, 2012), one-on-one résumé reviews with career advisors, and employer résumé reviews. In this thesis, the focus was on the assessment of one-on-one résumé reviews.

In a career development office, résumé writing assessment serves two purposes: *improving student learning outcomes* and *program evaluation*. The first purpose is assessment for learning outcomes improvement, which has been defined as a "cyclical process of gathering, analyzing, and interpreting evidence about institutional, departmental, divisional, or agency effectiveness to improve student learning" (Bresciani, 2009; Upcraft & Schuh, 1996). In turn, assessment data can then be incorporated into a body of evidence that serves the purpose of program evaluation, or "any effort to use assessment evidence to improve institutional, departmental, divisional, or agency effectiveness" (Upcraft & Schuh, 1996, p.19). In this thesis, 'evaluation' and 'assessment' will be used interchangeably to underline the multiple benefits of evaluating program goals. That is, the Résumé Review Appointment program is evaluated by assessing student learning outcomes as measured by a résumé writing rubric.

**The Résumé Review Appointment.** A résumé review program is the focus of this study. Similar to many college campuses, James Madison University (JMU) developed a program whereby students receive one-on-one résumé writing instruction from a career advisor. Typically, the programming starts when a student requests an appointment with the career office. Subsequently, students discuss various résumé aspects with the career advisor, after which they are encouraged to come back for a follow-up appointment. The number of follow-up appointments is unlimited, providing

opportunities for tracking skill development. That is, because students return with revised résumés, writing performance improvement can be captured during each new appointment.

Career advisors sought to investigate whether students improved in the various areas of résumé writing due to the résumé review appointments. To address the question of instructional impact on student learning and development, the career office created and followed an assessment plan with the assistance of the university assessment center. The center assisted with the development of a performance assessment rubric and a data collection design.

**The Résumé Ruler.** The Résumé Ruler was developed to measure student learning outcomes associated with participating in résumé writing appointments. Résumé writing student learning outcomes are specific objectives defining the skills and abilities students gain through résumé instruction and revision. The Résumé Ruler consists of two parts: a checklist portion and an analytic portion.

The checklist portion contains 25 items that can be categorized into five major areas: contact information, education, spelling/grammar, supplemental materials, and consistency. Consistency items include uses of font, punctuation, and format throughout the document. The Checklist portion of the Résumé Ruler was designed to account for the presence of essential résumé qualities.

The analytic portion of the Résumé Ruler reflects six essential résumé elements: Objective, Related Experience, Supporting/Secondary Experience, Organization, Headings, and Appearance. The analytic components are rated on a 1 to 4 scale supplemented with detailed behavioral anchors. Both portions of the measure were

revised continually to reflect the most current professional practice. The Résumé Ruler can be found in Appendix A.

**Performance Assessment Score Inference Validation**

The Résumé Ruler is a measurement tool used in a performance (or more accurately in this case "product," Fitzpatrick & Morrison, 1971) assessment. The alternative, contemporarily more common way of finding out whether a person can perform a task is asking her to choose a correct response from a number of options. This latter mode of measuring performance is "objective" in that subjective human error typically involved in performance quality judgment is absent (Cobb, 1998). In part because humans are absent from the scoring process, objective tests have practical advantages and have become a popular method of inquiry into unobservable traits (Messick, 1994; Shepard, 2000). However, they tend to provide indirect, incomplete indicators of ability. In many situations a more direct assessment is desirable (Black, 1998). Messick (1994) argues that performance assessments offset two major threats to validity of test score inferences: construct underrepresentation and construct-irrelevant variance. The downside of performance assessment is that human judgment and the use of a standardized rubric constitute additional sources of measurement error that are referred to as 'facets' in generalizability theory (Eckes, 2009).

A facet is a factor, variable, or component of the measurement situation that affects test scores in a systematic way (Eckes, 2009; Linacre & Wright, 2002; Wolfe & Dobria, 2008). Such variables can affect the difference between the unobservable true ability and an observed score. In an "objective" testing situation, frequently identified facets are the examinee facet and an item facet. The examinee's ability interacts with the difficulty of an item to produce an observed score. In a performance assessment situation

(or 'rater-mediated assessment'), additional facets are present – the *rater* facet and *scoring criteria* facet. Raters represent an unwanted source of potential error variation in observed scores that can threaten the validity of inferences drawn from assessment results (Eckes, 2009). Scoring criteria must be validated also to constitute valid and fair standards to which all observed performance is upheld.

Similarly, an examinee's observed score on a measure can be regarded as the sum of true score, measurement error, and random error (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach & Shavelson, 2004; Peter, 1979). Namely, error is the difference between an observed score and a true value of some unobservable trait. The presence of systematic and random measurement error undermines the validity of inferences made from performance scores. Measurement error is potentially amplified in performance assessments because additional sources of measurement error related to subjective scoring of complex responses emerge (Mehrens, 1992; Messick, 1994). Identifying and estimating measurement error magnitude can assist measurement instrument refinement. For example, if a large proportion of rating variability is found to arise from measurement error, and not from true score variability, the question becomes, "how can we make this assessment better?" Fortunately, common sources of error are well known, and often include rater biases and rating scale inconsistencies (Engelhard, 2002; McNamara, 2000; Mehrens, 1992). Being able to define sources of error variability allows one to estimate the magnitude of variability stemming from each source. Once serious systematic measurement error issues are identified, assessment practitioners can "go to the source" of the measurement errors and wield better control over them. The development of reliable, valid, and fair measures of résumé writing

hinges on the application of well-designed methods to deal with multiple sources of variability characterizing performance assessments.

Solutions to rater unreliability include making improvements to rater training (e.g., Ivancevich, 1979; Lumley & McNamara, 1995) or clarifying the rubric used to score artifacts (e.g., Kingstrom & Bass, 1981). Because résumé appointment assessment is used for learning assessment and program evaluation, identifying and eliminating sources of error is paramount. Previous assessment studies using earlier versions of the Résumé Ruler revealed gains in résumé writing quality from first to last appointments; however, the error associated with the estimates of means was substantial. For example, when two raters reviewed ten résumés in teams of two, the exact rater agreement was 62 percent. Exact rater agreement is the percentage of ratings between two raters that is exactly the same (e.g., Rater one and rater two both assign a '3' to the Objective criterion of the rubric). Thus of all possible scores on the analytic rubric, a rater team provided the same rating 62 percent of the time. An additional 19 percent of the ratings were no more than one rating apart. Considering there were only four possible scores on the rubric, both exact and adjacent agreement appeared low. Thus, even though individual students were not impacted by the ratings, systematic rater inconsistency can inflate or deflate the true programmatic impact on student learning.

In line with acting on assessment results, in an attempt to improve the precision of student ability estimates, both the rater training process and the Résumé Ruler rubric have been revised. Thus, over the last few years the career office has not obtained program evaluation evidence, electing to focus on instrument development instead. Therefore, it is important to reexamine the internal structure of the rubric scores to ascertain whether

the modifications enhanced score precision. Moreover, although résumé rubrics are prominent on college campuses, this researcher could not find psychometric evidence for any of them. Further development of this instrument based on empirical results can potentially result in a standardized instrument that could be used across campuses. In this thesis, rater effects and rubric elements are analyzed to provide recommendations for further development of the Résumé Ruler and the rating process.

**Structuring Validity Evidence Collection**

One purpose of this study is to gather construct validity evidence for performance assessment rubric scores used to evaluate student résumés in a career planning office. Kane, Crook, and Cohen (1999) offer types of evidence that can be used to inform the validity regarding performance assessment score use and interpretation. Under this framework, supporting evidence was presented for a series of assumptions regarding test score use and interpretation inferences. The resulting *interpretive argument* for the Résumé Rubric scores is reinforced by evidence for a selected set of statements that underlie inferences made from rubric scores. The presentation of an argument for performance score interpretations can manifest in particular successive stages. Although at least six distinct stages have been described in the literature (Chapelle, Enright & Jamieson, 2010), the focus of the résumé review literature is limited to assumptions underlying the domain description, evaluation, and generalization inferences. To this end, existing research on résumé content, rubric element analysis, and introduction to classical and generalizability theory provide support for such inferences.

In summary, when researchers make inferences from observations to unobservable traits such as students' résumé writing ability, they are crossing a succession of "bridges" from the operationalized, tangible ability indicators to the

description of latent traits. In making the interpretive argument for the Résumé Ruler rubric scores, assumptions are presented for three inferences: domain description, evaluation, and generalization (Kane et al., 1999). Investigation under the subsequent inferences will be possible depending on the results of the generalization inference validation. The inferences of explanation, extrapolation, and utilization are also discussed briefly in the next section.

**Inferences about Performance Assessment Scores**

In order to interpret scores on some measure of domain competence, the domain must be carefully described and reflected in the measurement tool (Kane, 2004). One assumption that underlies the *domain description* inference is *that it is possible to identify important skills needed for creating professional résumés.* Another assumption is that *simulation of important skills during an assessment task is possible.* The simulated skills must be necessarily important to situations in which the inferred performance occurs. In the context of the résumé writing quality domain, important skills have been identified by professionals and research. Skill simulation is also possible given that students produce the product to be used in actual employment evaluation situations.

Once the domain has been adequately described, the *evaluation inference* connects student performance to observed scores. This inference is supported by the assumption that *the performance scoring criteria are appropriate and have been applied as intended* (Kane et al., 1999). Subsequent to examining and confirming the fidelity of criteria appropriateness, the next step is to examine the logic behind possible *generalization* of the observed scores (Kane et al., 1999, p. 10). Obviously, one cannot observe scores on all possible student-written résumés; however, those hypothetical

résumé scores are inferred based on the observation in the performance assessment. To the extent that an observed score can be generalized, it is possible to make valid inferences about all possible performances in the résumé writing sample domain. One assumption regarding generalizations of Résumé Ruler scores is that *a sufficient number of rubric elements or raters was included to constitute cogent estimates about overall résumé writing ability*. A more technical formulation of this assumption is offered in the literature review drawing from a better understanding of rater effects and generalizability theory. Appropriately, generalizability theory was utilized to examine this assumption because it allows the identification of effects of raters and scoring criteria, facets that can weaken researchers' ability to generalize from observed scores (Kane, 1992).

While not specifically investigated in this thesis, the last three inferences could be examined at a more advanced stage of rubric validation. This is needed because reliability and generalizability are necessary but not sufficient to defend the meaningfulness of scores. The fourth stage of the inference validation process is *explanation,* an inference that requires evidence toward a salient resemblance between the empirical domain and the theoretical domain. The explanation inference is relevant because expected scores are assumed to represent a construct of résumé writing ability. Because a construct (i.e., résumé writing ability) is used in score interpretation, an explanation inference needs to be supported. Assumptions associated with this inference may be 1) that the résumé content knowledge and writing skills required to create a professional résumé vary across positions to which one may apply; 2) performance on the Résumé Ruler relates to performance on other measures of résumé writing quality (e.g., employment status); and 3) the internal structure on the Résumé Ruler ratings is

consistent with theory as a number of interrelated elements. Common variance among elements on a rubric constitutes a construct that is présuméd to drive or *explain* scores manifested through performance assessments. Kane (2001) asserts that the explanation inference is relevant when the construct is used in score interpretation.

The *extrapolation* inference is relevant because the construct of résumé writing quality assessed using the Résumé Ruler accounts for the level of proficiency that would be perceived in a professional setting. Because résumé writing performance is related to other contexts where résumé quality is important (i.e., outside of a career counselor's office) extrapolation is a relevant inference. The extrapolation inference includes an assumption that the construct of résumé writing quality assessed using the Résumé Ruler accounts for the level of proficiency that would be perceived in a professional setting. That is, the rubric quality scores represent quality in the target domain (Kane et al., 1999) of professional résumé review.

Finally, the *utilization* inference makes the connection between the target (i.e., the expected) Résumé Ruler score and student-related decisions. Utilization inferences strengthen the arguments for decisions made based on a test score (Bachman, 2005). An assumption underlying the utilization inference is *that Student Affairs administrators can interpret the meaning of the scores and that improvement to individual instruction can be made based on the résumé ratings*. Because this type of assessment is low-stakes, this inference is not examined here, but should be considered as it pertains to supporting the validity argument. In fact, although decisions about Résumé Ruler scores do not affect the students immediately, the scores do contribute to the distribution and efficiency of resources allotted to career-related programs.

To summarize, the assumptions that observations represent test scores, and test scores represent student ability must be scrutinized (Kane et al., 1999; Kane 2001; 2004). The six aforementioned inferences enable stakeholders to better evaluate the meaningfulness of résumé writing scores. In this thesis, research on specific elements of the rubric was presented to provide the evidence for the domain description inference. These studies help justify or discredit the use of particular elements in the rubric. A conclusion that performance observations are relevant allows the examination of the next inference – evaluation. The evidence for the evaluation inference includes the description of the content expert rubric development process and rater effects on assessment scores. At this stage, studies of element analysis can be presented (Chapelle, Enright, & Jamieson, 2008). The conclusion that the observed score is appropriate allows continuation to the next inference. During the generalization stage, Chapelle et al. (2010) recommend conducting generalizability and reliability studies as well as scaling and equating studies. Further, support for generalization to the expected performance can be made from standardizing task administration conditions. In the case of the Résumé Ruler, this is not applicable because there is only one task (i.e., one written résumé), and because students are free to complete the task at their own pace in a self-selected setting. In the case of the Résumé Ruler, generalizability support can lead to the conclusion that the observed scores reflect the expected scores across raters. Thus, the focus of this study is on the first three stages of the inference-making process in the interpretive argument for Résumé Ruler scores.

CHAPTER TWO

# Literature Review

This thesis investigates the assessment of student résumés by a career services office. Specifically, it explores systematic errors in performance assessment ratings of student résumés using the Résumé Ruler rubric. Such evidence adds to the measure's validity argument; and specifically to the validity of inferences regarding the generalizability of student performance assessment scores (Kane, Crook & Cohen, 1999). Ultimately, the concern lies with the inferences made regarding student learning and development related to résumé appointment instruction conducted on a college campus. To begin, the literature review overviews current trends in Student Affairs assessment, continues with the functions of career offices, the assessment of résumé writing, and finally focuses on methodology using Kane's (1999) validity framework. The methodology review initially addresses performance assessment (PA) and rubrics, narrowing to specific indicators produced by generalizability analyses that address the dependability of PA ratings.

## Current Trends in Student Affairs Assessment

According to Ewell (2002) the general Assessment Movement in higher education was spurred in late 1970s by the increased pressure from states and demands from "consumers of education" for accountability. Specifically, these stakeholders wanted to know how much students were learning as a result of the college experience, which is paid for through tuition and, often, state and federal support. At the time, the emphasis was primarily on academic programs. Student affairs (SA) followed into the assessment arena more recently (Schuh, 2009). Indeed, in today's environment of limited resources, the accountability pressures for all units – regardless of division – are heavy.

To help SA units respond to accountability pressures, several authors and organizations have published assessment guides tailored toward SA. *Learning Reconsidered* (National Association of Student Personnel Administrators (NASPA) & American College Personnel Association (ACPA), 2004), *Learning Reconsidered 2* (Keeling, 2006), *ACPA ASK Standards* (ACPA, 2006), and *Assessment Reconsidered* (Keeling, Wall, Underhile & Dungy, 2008) represent comprehensive recommendations for systematizing improvements of student learning in higher education. Integrating SA professional standards and these seminal publications, Shutt et al. (2012) proposed a model of best practices in SA assessment. According to this model, programs must first make the *decision to commit to intentionality*, which implies organized institutional support of staff time toward systematic program improvements. At the second stage, programs should focus on *an outcome,* which implies that a set of predetermined outcomes is benchmarked. During the third stage, SA offices establish mechanisms for *assessing the outcomes.* Finally, a program reaches its full efficiency potential by becoming *peer-reviewed.* This includes external validation processes like sharing practices and effects with other universities through conference presentations and peer-reviewed manuscripts. The model is designed to continually loop, where reflection and feedback from external constituents shape subsequent iterations.

The model proposed by Shutt and colleagues is rooted in principles similar to those used by academic programs. For example, the key features - predetermined student learning outcomes, evaluation of results (i.e., assessing outcomes), using assessment results and dissemination (e.g., external review) - have been consistently endorsed by assessment experts (e.g., Bresciani, 2009; Erwin, 1991; Palomba & Banta, 1999; Suskie,

2009). Also, like the models proposed for academic programs and institutions, the framework is flexible enough to accommodate a diverse set of programs in student affairs.

## Assessing Résumé Writing

SA student learning outcomes assessment follows the cycle of (1) developing specific learning outcomes and relating them to programming, (2) evaluating the collected data, (3) providing results, (4) using the results for improvement of programming and assessment and (5) disseminating the results for peer-review (Bresciani, 2009; Shutt et al., 2012). Using this framework, the assessment of résumé writing was introduced.

**Defining student learning outcomes.** Assessment researchers (e.g., Bresciani, 2009; Erwin, 1991; Palomba & Banta, 1999; Suskie, 2009) identify the development of student learning objectives as the first step in guiding assessment practice. Assessment practitioners should state student learning objectives that are student-oriented, meaningful, specific, manageable, and measurable (Bresciani, 2009). More importantly, the measurement of a learning objective must provide evidence that can contribute to continuous program improvement (Bresciani, 2009). Overall, the learning objectives must identify what students are expected to know or be able to do as a result of intervention (Suskie, 2006). Programming should be linked directly back to objectives to demonstrate the connection between expectations of students and provided resources to meet those expectations. A learning objective could be framed in terms of achievement of a standard, or change in amount of a trait over time. Whereas the former connotes that students must be proficient relative to a predetermined benchmark, the latter indicates that a certain amount of learning has occurred due to the course of instruction. In both

situations, some extent of causal relationship inferences between instruction and learning outcomes is typically made. The extent to which inferences of causality are valid is beyond the scope of this treatment, but interested readers should consult Nichols (2007). The verb choice describing the knowledge and skills expected of students as a result of a program are an important consideration toward objective development. Krathwohl (2002) suggests using the revised Bloom's taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956).

In the résumé review program example, the career office developed two outcomes: As a result of participating in résumé reviews students will: (1) "…demonstrate knowledge of the fundamental components of a well-written résumé"; and (2) "…*create* a well-written résumé." Fundamental components are implied to be résumé elements that are necessary to any résumé (e.g., Related Experiences). Also, it should be noted that although the official objective calls for "knowledge", the very nature of the assessment (i.e., to produce an artifact) requires higher order cognitive abilities that call for generating a product. Because cognitive processes build on each other in complexity (Krathwohl, 2002), the "knowledge of fundamental components" is nested within the cognitive skill "create" (Objective 2). The office further defined the "fundamental components" and "well-written" based on a literature review and subject-matter expertise. Over a dozen career advisors reviewed the components and criteria for a well-written résumé on several occasions, until unanimous agreement was achieved. If the two objectives are met at a group level, it is inferred that students "know fundamental components of a résumé" and that in general they can "create a well-written résumé."

**Selecting systematic methods for evaluating progress on objectives.** Having defined the expected student learning outcomes, the next step is choosing the specific measurement instrument for testing résumé writing skills (e.g., multiple choice test, essay tasks, portfolio). The choice of assessment method depends on the nature of a particular cognitive process targeted by the learning objective. For example, when assessing factual information recall, a multiple choice test could be a viable mode of measurement (Krathwohl, 2002). Although it is possible to measure more complex cognitive abilities using multiple choice items (e.g., Cobb, 1998), the fidelity between the assessment performance (i.e., choosing response) and real-life performance (e.g., developing a clear, succinct, but specific job objective) is lower. On the other hand, asking students to perform an assessment task they would have to do "in the field" produces more confidence about that skill level. In the case of résumé writing, a multiple choice test for knowledge would measure recall of information about various résumé elements. To evaluate learning, this test would be administered at multiple time points, necessarily before instruction and after.

However, defining outcomes simply in terms of knowledge about résumé composition neglects the context. Because writing a résumé involves *creating* an actual product, the logical choice in measuring students' ability to create effective résumés is to evaluate the "constructed response." Constructed response assessments are a type of performance assessment described in detail later. It is difficult to make an argument that identifying the correct essential elements on a résumé represents students' actual ability to create an effective résumé (Nitko, 1996; Osterlind, 1997). In fact, consistent with Bloom's et al., (1956) taxonomy, students are effectively *synthesizing* their self-

knowledge and facts about résumé construction by producing their own résumés. It can be argued that the execution of a professional résumé is a complex task akin to composing a writing sample. Using the latter rationale, performance assessment is the logical assessment approach.

In scoring résumés, career advisors have a choice between using checklists and rubrics in producing a score that represents students' mastery. Both checklists and rubrics clarify educator expectations to students (Groeber, 2006). Checklists are necessarily lists that evaluate the presence or absence of an element in student performance (Moskal & Leydens, 2000). On the other hand, the term "rubric" refers to a scoring guide used to evaluate the quality of students' constructed responses (Moksal & Leydens, 2000; Popham, 1997). Although both rubrics and checklists offer students more information than a letter or numerical grade, the rubric takes the criteria list one step further by associating stated performance standards with graduated levels of mastery. For example, a checklist might remind students that they must edit written reports for punctuation errors; however, a rubric states the same objective with graduated levels of mastery: 0–1 errors = 4 points, 2–4 errors = 3 points, and so on. This additional evaluative advantage allows educators to distinguish between minor lapses in punctuation (2 or 3 punctuation errors), and a significant lack of understanding of correct punctuation (8 to 10 errors). This characteristic makes rubrics a more nuanced evaluation instrument than a checklist.

The résumé review program under study created an assessment rubric that measured the quality of specific résumé elements. As an office, the career advisors conducted a search for existing résumé rubrics and found that most were too broad for

addressing writing skill development in several components of the résumé. An advisory committee was formed, and the staff members outlined the essential components of a résumé. The essential components were subsequently divided into two measurement tools – a checklist and a rubric. The checklist criteria included the presence or absence of certain information or qualities in a résumé (e.g., contact information, education information, and consistency). On the other hand, the analytic criteria included content characterized by varying quality levels (e.g., objective, articulation of related experience, overall appearance). The résumé rubric was evaluated and modified several times based on staff input.

**Measuring traits.** Program outcomes can be measured via three strategies: (1) the value-added approach, (2) the career success approach, and (3) the impact approach (Jennings & Lumpkin, 1989). In this study the value-added approach is utilized to examine the effect of instruction of program outcomes. The next step in an assessment cycle is to design and implement data collection (Bresciani, 2009), which involves selecting, sampling, and soliciting students and selecting instrumentation (Schuh, 2009). Before embarking on measuring learning outcomes, one should consider the inferences intended from the collected information.

Depending on whether the goal is to infer *learning* or *achievement*, different data collection strategies can be employed. Student *learning* has been defined as "a relatively permanent change in a person's behavior over time that is due to experience rather than maturation" (Shavelson, 2009, p.10). According to this definition, one can say that some *value* has been added to a student's experience. In order to detect value added to a student's experience, assessment practitioners must measure a quality at least twice,

before and after an implemented program. However, "learning" has been used interchangeably with "*achievement",* which is the amount of knowledge or skills at a specific time that a person has accumulated (Shavelson, 2009, p.11). In contrast to learning, achievement only needs to be measured once, hence the notion of total accumulation of knowledge. Given the definition of student learning, much of the current assessment practice does not differentiate between learning and achievement. Thus, it is important to emphasize that the appropriateness of making inferences about student learning depends on the type of measurement design. When objectives emphasize student learning, the measurement of a trait should occur more than once to account for this added value. Longitudinal measurement allows practitioners to infer change due to an intervention.

Given that the unit in question wished to know how much students had learned as a function of its programming, student performance was measured at two time points. Specifically, because students could schedule several appointments, career advisors could evaluate the progression of résumé writing skill development over time. It should be noted that because students could set up an unlimited number of follow-up appointments, post-résumés represented any number of reviewed draft iterations. Students set up appointments with any available advisor via calling or emailing the career development office. Students represented a variety of disciplines including education, business, social and health sciences, media arts, and engineering. Appointments were scheduled via an electronic scheduling system that allows the student to choose a preferred advisor. The advisors and students worked one-on-one for one hour, advisors providing verbal and written feedback to students regarding each résumé. Each student's initial résumé and

the last appointment résumé were scored during another designated time. Thus, advisors did not use either the checklist or the rubric to judge the quality of each résumé until the time of the study. Students were not referred to any materials other than advisor feedback in improving their work during each appointment. Checklist and rubric scores were not shared with students during this study; they were used for program evaluation purposes only. In this thesis, the value-added quantity of student's résumé writing improvement is measured after evaluating the dependability of scores at both the initial and final measurement points.

**Presenting results.** Assessment results should be disseminated to internal and external stakeholders (Bresciani, 2009; Ewell, 2002; Suskie, 2009). Shutt and colleagues (2012) describe this process as the assessment stage where the program becomes *peer-reviewed*. External validation occurs in the form of sharing practices and effects with other universities through conference presentations and peer-reviewed manuscripts, whereas internal validation can occur within the university by other stakeholders examining the importance of assessment results (Harvey & Knight, 1996). Résumé instruction assessment may inform other career offices of particular instructional and assessment techniques.

The career office under study created an assessment report that was submitted to the director of career planning, director for student affairs, and to the institutional assessment office. Thus, the résumé review program was subjected to peer review.

**Using results for improvement.** The purpose of program assessment results is to make changes to programs, courses, departments, or any number of areas that might improve future iterations of programming (Erwin, 1991). Shutt et al. (2012) describe

their best assessment practice model as focused on assessment *cycles*. Using assessment results for improvement is the impetus of the assessment process; it is the reason assessment practitioners conduct assessment. Whether the cycle showed "significant" effects or not, some information is to be gained by reviewing and interpreting the assessment results. Subsequently, the program should decide whether or not modifications to the program or assessment methods are necessary. Even when a program observes "positive" effects, it is important to seek ways to provide validity evidence behind the effects. In other words, there should be some assurance that the assessment process produces accurate, fair, and useful information (Suskie, 2006; 2009).

At this stage, the résumé review program has not used the results for instructional improvement. Before an office is comfortable making decisions about the program, they must first be certain that the assessment is of the quality to warrant programmatic decisions. In other words, before attempting to infer a program effect, it is necessary to ensure that the measurement instruments are precise and produce consistent results. Wiggins (1993) reports that "few in-house-designed tests and assessments meet the most basic standards for technical credibility, intellectual defensibility, coherence with system goals, and fairness to students" (p. 19). Although résumé review assessment is of low-stakes, its potential role in preparing students for differential professional writing situations is significant. To maximize the positive impact on students, programs should strive to employ the most appropriate methods of evaluating its strengths and weaknesses. So far, résumé review assessment has focused on evaluating the quality of the Résumé Ruler. The rater-agreement information indicated that certain elements of the rubric were interpreted differently by raters. Some efforts to improve the rating

process have already been made by introducing additional training. The process of evaluating two different rating methodologies resulted in several changes designed to eliminate bias in the measurement procedures. Specifically, at least two raters were recommended for evaluating résumé quality. Further, in the current rating procedure, raters are not assigned students' first and last drafts to avoid bias. However, although measures have been taken to reduce measurement error due to rater inconsistencies, it is important to examine measurement error empirically.

In sum, the résumé review program has been following assessment practices reflective of expert advice. Résumé writing assessment utilizes performance assessment, a measurement method that can introduce observed score errors beyond those of direct measures (Kane, Cooks & Cohen, 1999; Mehrens, 1992). Performance assessment has been viewed as "a sample of student performance drawn from a complex universe defined by a combination of possible [performance] tasks, occasions, raters, and measurement methods" (Shavelson, Baxter & Gao, 1993). Therefore, assigned scores can vary from one rater to another or from one occasion to another. This type of variability is *measurement error due to sampling variability* (Shavelson et al., 1993), and affects the inferences we can make from student scores. Currently, the key need for examination is the variability of the collected résumé rubric scores due to raters. Given that résumé scores are manifestations of rater observations, it is of utmost importance to examine rater effects. Previous research on the magnitude and sources of bias in performance assessment informs the measurement design in this study.

**Performance Assessment**

Performance assessment (PA) provides rich information about skills and knowledge closely related to real-world activities. Given this attractive characteristic, PA is often utilized for evaluation of complex student learning and development outcomes. Indeed, PA and use of rubrics has become popular in high-stakes testing. Ratings of student performance on Advanced Placement English examinations (Braun, 1988; Wolfe, 2009), tests of language knowledge (Eckes, 2009; Lumley & McNamara, 1995; Myford & Wolfe, 2002), state writing tests (Engelhard, 1994; Du & Wright, 1997), international university writing tests (Farrokhi & Esfandiari, 2011; Sudweeks, Reeve, & Bradshaw, 2004) and certification exams (Lunz, Wright & Linacre, 1990; Smith et al., 1994) are only examples of recent high stakes PAs.

Given the increased PA use, it is important to be as clear as possible about the intended interpretations and uses of PA scores. Further, clarity about the type of evidence needed to validate these interpretations and uses (Kane et al., 1999) is essential to efficiency and accuracy in the interpretation process. According to the Educational Testing Service (ETS) PA is

> "…a test in which the test taker actually demonstrates the skills the test is intended to measure by doing real-world tasks that require those skills, rather than by answering questions asking how to do them. Typically, those tasks involve actions other than marking a space on an answer sheet or clicking a button on a computer screen. A pencil-and-paper test can be a performance assessment, but only if the skills to be measured can be done, in a real-world context, with a pencil and paper" (ETS, 2013).

Kane, Crooks and Cohen (1999) agree with ETS's view on the differences between PA and objective measures. For PA they identify the ability to measure complex cognitive skills, real-world performance to task-performance match, and instructional value as benefits. On the other hand, scoring inconsistencies, low correlation between task scores, and difficulty in equating different forms of PAs are highlighted limitations.

**Advantages.** Advantages associated with PAs make them attractive tools in extracting meaning from student ability indicators. Although the measurement of complex cognitive skills is possible using objective measures (Cobb, 1998), it can be time and labor intensive. At times it is more appropriate to identify a skill desired in a student and ask them to demonstrate it (Kane et al., 1999). PA also offers a more straightforward inference line to the domain of performance. When discussing real-world performance to task-performance match, Kane and colleagues (1999) refer to the fidelity of performance task-to-outcomes relation. For example, having provided the reasoning behind choosing a particular experimental design, a student has supplied a sample of real-life performance by providing a rationale for selecting a particular experimental design.

Further, PA has influenced student learning and teacher instructional approaches (Groeber, 2006). When using PAs, the goal is to score a task that is similar to that outside of assessment, and students are required to think about what a performance task entails; thus it is required that instructors and students think about real-life performance. Furthermore, PAs allow for evaluation of the entire process required for completion of a product, which offers a stronger basis for score validity (Wiggins, 1989). PAs are intended to "narrow the gap between the observed performance and the proposed

interpretation by basing the assessment on samples of the kinds of performance referenced in the interpretation" (Kane et al., p. 7).

**Disadvantages.** Nevertheless, measuring cognitive skills with PA is not a simple task. Performance-based assessments have been criticized for the additional cost, time, the necessity for subjective human judgment, and the resulting lack of reliability and validity (Shavelson, Baxter & Zou, 1993). The "independence" of tasks, rater effects, and difficulty in equating different PA forms are a few glaring issues (Kane, Crooks & Cohen, 1999). Although not an exhaustive discussion of measurement error in performance assessment situations, the following section discusses several prominent measurement issues encountered in performance assessment. Specifically, because PA requires the subjective judgment of raters, rubrics constitute a useful tool in helping raters attend to the complex task performances. With the use of rubrics come several issues related to rater inconsistencies and rubric element dependencies.

## The Process of Gathering Validity Evidence

While dissenting opinions about validity still persist (e.g., Borsboom et al., 2004) the educational measurement field leans toward conceptualizing validity as the degree to which arguments support the interpretations and uses of test scores (Kane, 1992; 2006; Messick, 1989). Kane's interpretive argument (1992) is one such process of score validation that explicitly calls for consideration of test interpretation and use consequences. In this thesis, an interpretive argument guides the evaluation of a résumé writing rubric. It should be noted that the components of an interpretive argument vary with the context of the testing situation. Thus, the focus here is on general statements pertinent to performance assessment validation.

Interpretation validity assigned to PA scores depends on the plausibility of the inferences involved in the interpretation (Kane, Crooks, & Cohen, 1999). The sequence of inferences from observed performance to score interpretations and uses constitutes an *interpretive argument* (Kane, 1992; 2004; 2006). As with any measurement technique, the goal of PA is to make inferences from a sample of performances to a domain of performances. The difference between PA and objective tests is that the assessed performance is more likely to resemble the type of performance conducted in the target domain (Kane et al., 1999).

Kane and colleagues (1999) describe domains to which researchers make inferences about PA. Ultimately, PA score interpretation involves inferences from the observed performance to a wide domain of performances called the *target domain.* Kane et al., (1999) postulate that target domains in education are very broad. For example, the domain "critical thinking" may include a wide range of tasks in many contexts, ranging from writing an email to composing a thesis. The professional writing domain is less broad, however; with appropriate evidence one can argue professional writing tasks subsume résumé writing tasks. The second assumption about the definition of PA is that PA involves a sample of performances from the target domain. For instance, does the task that elicits student writing performance in the 'past professional experience' section of a résumé also exist in the target domain of writing in general? When a set of observed performances is thought to be a random or representative sample it is known as a sub-domain of the *universe of generalization*.

Assuming that the universe of generalization is a subset of the target domain, Kane and colleagues (1999) identify at least three critical links to the chain of inferences

from the observed performance to the expected performance averaged over the target domain. First, raters score student performance, which yields an observed score. Second, researchers can generalize the observed score to the universe score, defined over the universe of generalization. Third, the universe score is then extrapolated to the target score, defined over the target domain. Chapelle, Enright, and Jamieson (2010) also emphasize the importance of describing the domain of interest, explanation inferences, and score utilization inferences in performance assessment. Evidence supporting the cogency of ratings and ruling out alternative explanations for observed effects involves a critical review of the scoring rubrics, the scoring procedures, and the procedures for administering the assessment (Chapelle et al., 2010). The conclusion that observations are relevant allows the examination of the next inference (Kane et al., 1999). Next, assumptions associated with each step in the résumé rubric interpretive argument are listed and validity evidence is presented.

**Domain description.** Kane, Crook and Cohen (1999) also assert that two assumptions are embedded in the definition of performance assessment itself. First, the definition assumes that *the interpretation of examinee scores will emphasize levels of skill in some performance domain*. To this end, research on specific elements of the rubric is presented. These studies bolster the inclusion of certain elements on the rubric deemed important by human resource staff and administrators. Second, *the observations used to draw inferences about skill in this domain involve performances on tasks from the domain of interest.* The second assumption is supported by the fact that the product (i. e., résumé draft) is the same that would be composed for the domain of interest.

***Assumption 1.*** *The observations used to draw inferences about skill in the résumé writing domain involve performances on tasks from the domain of interest.* Important skills needed for creating professional résumés can be identified using the Résumé Ruler. Numerous studies have focused on manipulating various content features of résumés in examining effects on readers' judgments (e.g., Field & Holley, 1974). In general, most job applicants must be able to maintain flexibility in résumé content "through life-long learning and adapting to transitions" (Savickas, American Psychological Association, 2010). Students must learn to modify their résumés without compromising quality. It is likely that college graduates will hold more than one job position in their lifetime, and the positions may vary in nature to a certain extent. Ross and Young (2005) emphasize that résumé content information is necessary to be presented in different ways depending on which occupation one is seeking. They listed job or career objective, education, and work experience as key elements. Focusing on construct representativeness as one of the assumptions in making inferences about observations and observed scores, the domain of quality résumé content is presented next.

*Objective.* The objective statement on a résumé is a statement that connotes job applicant goals and objectives regarding the particular job position or career. A professional objective appears to be an important component to hiring officers (Hornby & Smith, 1995; Hutchinson, 1984; Hutchinson & Brefka, 1997). Hornsby and Smith (1995) recommended that the objective should be included at the beginning of the résumé, and should relate to the needs and goals of the organization to which one is applying. Further, professional objectives should be specific; applicants who are seeking different types of jobs should tailor the objective and résumé content to each particular position.

However, Ross and Young (2005) argue that objectives are only somewhat important, and variation in perceptions of importance is affected by the job discipline. In general, research on résumé objectives suggests that objectives serve as an important component of a well-written résumé.

*Related experience.* One of the most important elements of the résumé is related experience (Hornby & Smith, 1995; Hutchinson, 1984; Hutchinson & Brefka, 1997; Ross & Young, 2005). Related experience includes the employment history - employment dates and company addresses, internships related to the applicant's job of interest, and volunteer experience (Hornsby & Smith, 1995). This essential information requires careful and specific articulation of previous job duties and acquired or developed skills. Importantly, these facts should not stand alone, but demonstrate students' ability to select relevant and appropriately detailed information that best aligns with a prospective position (Charney, Rayman, & Ferreira-Buckley, 1992). Klemp (1977) identified the ability to organize information as one of the most important to professional success. Tsai, Chi, Huang, and Hsu (2011) found that applicant work experience and educational background increased recruiter hiring recommendations, whereas applicant work experience predicted recruiter perceived person-organization fit. Charney et al. (1992) found that résumé readers gave significantly higher ratings to résumés that had highly relevant experiences listed in the "related experiences" section than moderately relevant experiences.

*Supporting/Secondary Experience.* Supporting, or general information should follow the related experience information (Hutchinson, 1984). Secondary information includes awards and extracurricular activity. Hornby and Smith (1995) concluded that

human resource professionals expect information about all applicant work experience, not just the experience related to the specific job. They found strong preference among human resource professionals for résumé items that document honors and awards. Generally, it is accepted that all achievements from both college and work experience are present on a résumé, including scholarship, student government, athletic involvement, and community experiences. If the space on the résumé document is limited, the achievements that connote leadership, communication, organization, and collaboration are essential items for inclusion (Hornsby & Smith, 1995).

*Résumé organization, headings, and appearance.* The last three components of the Résumé Ruler diverge from focusing on content, and instead focus on résumé presentation. Schramm & Dortch (1991) found that many typographical and grammatical errors, word choice, spelling, photocopy reproduction, and length of the résumé errors may cause employers to lose interest in the candidate. Charney et al., (1992) found that recruiters rate error-free résumés significantly higher. This suggests that when constructing a résumé issues in appearance and format may be weighted as heavily as the issues of content. Good organization, neat appearance, and length of not more than one to two pages are important (Harcourt & Krizan, 1989; Pibal, 1985, Schram & Dortch, 1991). Employers prefer bullets, boldface type headings, and underlining. Human resource professionals prefer the chronological, list format of résumé data presentation (Schramm & Dortch, 1991, Stanley-Weigand, 1991). This is suggested for both new and experienced job applicants because gaps in work history are most obvious using such organization style. One of Girrell's (1979) Axioms states that every concept in the résumé should be rank-ordered from most important to least important as the résumé

progresses, but also within each essential element in the résumé. The majority of more recent sources (Adams & Morin, 1999; Bortoli, 1997; Brown & Hayes, 1998; Lovelace, 2001; Nichols, 2001) suggested that résumés absolutely should be no longer than one page in length. Researchers' empirical attention to résumé organization and appearance suggest a matched professional attention of administrators and human resource managers. McDowell (1987) studied mechanics and order. The results of the study suggested that résumé reviewers such as human resource managers consider grammar and logical, orderly organization equally as important in a quality résumé. Ease of access to pertinent information is paramount as most reviewers in business settings spend between 5 and 45 seconds reading each résumé (Lovelace, 2001).

**Evaluation/Scoring.** The evidence for the evaluation inference can include the description of the content expert rubric development process. At this stage, studies of element analysis can be presented (Chapelle et al., 2008). Kane et al. (1999) describe the appropriateness of making inferences from performances to observed scores as dependent on one relevant assumption: the criteria used to score the performance must be appropriate and have been applied as intended. The criteria assumption includes considerations of the scoring rubrics criteria appropriateness important to preventing construct irrelevant variance (Chapelle, 2010; Kane, 1992). In contrast, failure to include some pertinent criteria can lead to construct underrepresentation. Conditions under which score interpretation is intended may include rater selection and training or clear communication of the rubric criteria nature. Rubric development procedures used in the construction of the Résumé Ruler were described below.

***Assumption 2.*** *The criteria used to score the performance are appropriate and have been applied as intended*. Scoring criteria present an additional source of measurement error over and above that of objective tests. To this end, it is important to review the characteristics of a résumé.

*Rubric construction*. Rubrics can be designed to be holistic or analytic (Popham, 1997). Whereas a holistic rubric requires raters to score the process or product as a whole without judging the component parts separately, an analytic rubric requires raters to score specific parts of the performance and then to sum specific scores (Moskal & Leydens, 2000; Nitko, 2001). At most, limited feedback is provided to the student when scoring performance tasks in this manner. With analytic rubrics, students receive specific feedback on their performance with respect to each of the individual scoring criteria (Nitko, 2001).

Evaluative criteria are used to distinguish acceptable responses from unacceptable responses and can be weighted equally or differently. Quality definitions describe the way that qualitative differences in students' responses are to be judged. For example, in a writing situation “mechanics” and “style” are popular evaluative criteria. For each qualitative level, a description must be present defining the range of possible performance.

Quality definitions for each criterion are also called behavioral descriptions or anchors. The recommended number of such levels is four to five (Popham, 1997). Usually, these levels are associated with a quantitative label to facilitate quantitative analyses of performance and a general label associated with that performance (e.g., advanced). Each behavioral description must be consistent with criteria according to the

common verbiage. For example, the "usage and mechanics" criterion in the JMU Writing Rubric defines "beginning" as containing "pervasive errors in mechanics, usage, grammar, or sentence structure" whereas "developing" is described as containing "some errors in mechanics, usage, grammar, or sentence structure." It is important to note that these are parallel behavior anchors in that only one word is manipulated in describing the extent to which a certain criterion is met. Although systematizing behavioral anchors can eliminate some biases, it can also subtract from the richness of performance description.

In résumé writing PA, each rubric criterion represents student résumé writing performance on each résumé element. The score on each task may be affected by the performance on another task. Dependence between criteria may be introduced by criteria being somewhat inclusive of each other. For example, a résumé rubric may contain the criterion "quality of headings" and also the criterion "quality of organization." It could be that the particular way that a student used her headings helped organize the contents of the résumé. It is important to evaluate score precision associated with each element. That is, similar ratings on some rubric elements may be more difficult to achieve than on other elements; the extent to which raters respond to criteria differentially is a serious source of unwanted variability.

*General rater effects.* Raters evaluate student performance on an evaluative rubric. The appropriateness of use depends partially on the level of possible rater effect present. Rater effects studies typically focus on specific domains - rater cognition, rater characteristics, tasks and environment, and development of statistical modeling techniques to correct for rater effects (Wolfe, 2004). Because this thesis is concerned with identifying rater effects, rater biases, and their manifestation in PA contexts were

described. Saal, Downey, and Lahey (1980) outlined four major categories of rater errors: 1) severity 2) halo, 3) central tendency, and 4) restriction of range. Rater accuracy may be affected by several factors (Engelhard, 1994; Myford & Wolfe, 2002) - rater experience, cognitive factors, and characteristics of the rating criteria. Rater inaccuracy results in low levels of consistency between assigned ratings and expected ratings. Due to various factors such as inadequate training, misalignment with the worldview of the rubric or distractions, judges may provide unexpected ratings given true examinee ability (Myford & Wolfe, 2002; 2004). To counterbalance this problem, it is best practice to use as many raters as possible when producing PA scores (Myford & Wolfe, 2002). In classical test theory, a *true score* is defined as the average of all possible examinee scores over an infinite number of test-taking occasions (and raters). Given this definition of true score, one could argue that the average rating across several raters likely reflects the true score of the person being assessed better than a single evaluator's rating. In generalizability theory – which is described in more detail later – one would say that scores based on more raters typically better represent a person's *universe* score.

*Rater harshness.* Rater severity or leniency (or rater harshness) is a rater's tendency to consistently provide ratings that are lower or higher than is warranted by student performances (Saal, Downey, & Lahey, 1980). Rater harshness occurs when raters use the same criteria, but prescribe scores with different levels of stringency (Myford & Wolfe, 2002). Literally, it means that one rater's mean score is different from another rater's mean score given the same performance task completed by a set of examinees. A situation that describes rater harshness may be represented by highly consistent rank ordering of scores between raters, but inconsistency of ratings between

raters relative to the scale of scores. For example, it is possible for two raters to rank-order a set of résumés by quality in the same exact way, with one rater consistently rating all résumés lower than the other rater. The term "harshness" can be used to describe overall rater severity and differences between rater interpretations of rating scale thresholds (Lumley & McNamara, 1995). If raters are consistent, then it may be possible to calibrate raters in ways that are similar to the calibration of test items and to adjust estimates of student competence for differences in rater severity (Engelhard, 1994). In practical terms, this – rater calibration - could be accomplished by including a common set of student compositions that are rated by several raters.

*Rater centrality/extremism.* Rater centrality represents raters' tendency to assign scores closer to the middle of the performance scale regardless of whether they have mastered the nature of the rating criteria (Engelhard, 1994; Myford & Wolfe, 2004; Saal, Downey & Lahey, 1980). This bias results in ratings in the middle of the scale regardless of examinee performance. Scores concentrated in the middle of the scale will exhibit low variability, which is associated with low reliability. This phenomenon is known as restriction of range; the effect introduces artificial dependency in ratings (Saal, Downey & Lahey, 1980). In contrast, rater extremism occurs when raters excessively use the extreme rating scale points regardless of examinee performance (Myford & Wolfe, 2002). Centrality is associated with large and positive residuals for low expected ratings, and large and negative residuals for high expected ratings. Conversely, rater extremism results in residuals that are near zero for extreme predicted ratings and the absolute value of residuals increases as the predicted ratings approach the center of score distribution (Wolfe, 2004).

*Restriction of range.* Restriction of range is "the extent to which obtained ratings discriminate among different ratees in terms of their respective performance levels" (Saal, Downey & Lahey, 1980). This bias attenuates researchers' ability to address whether or not the assessment process identified true individual performance differences. Restriction of range creates issues in conducting generalizability studies because when true score variability is restricted, identified rater or item biases can appear inflated in comparison. Although there are methods to identify and correct for this bias when large samples are available (Linacre, 1989), restriction of range obscures true proportions of true variance to measurement variance within the generalizability theory framework.

*Halo effects.* Engelhard (2002) describes halo effects as another type of important rater characteristic. This type of rater bias occurs when raters do not discriminate between conceptually dissimilar and independent aspects of examinee performance. For example, a résumé rubric may contain elements such as 'objective', 'appearance', 'relevant experiences', and 'skills'. A halo effect occurs when raters focus on one of the elements as the deciding factor of ratings across the entire PA, or make a holistic judgment. Similar to centrality, the rater error introduces artificial dependency in ratings. Residuals in this case would be random. For instance, if a rater was impressed by the student' ability to articulate a career objective, it is possible that the rater would then inflate ratings of other résumé writing rubric elements. In practice, halo effects become apparent when one rater uses a uniform rating pattern (e.g., "1111" or "4444") (Engelhard, 1994). Because the "source of halo" or the element that was particularly impressive is random, the discrepancy between expected and actual ratings will also be random. Halo effects obscure an examinee's true score (Farokhi & Esfandiari, 2011),

threatening the validity of inferences we can make from assessment results. Halo effects can be "true" or "illusory" (Murphy & Cleveland, 1991). True halo effects are not rater effects because a student may actually perform at the same level across all elements of a rubric or task. Similar to general rater centrality and restriction of range biases, halo effects contribute to obscuring the proportion of true score variability to measurement variability.

In summary, rater subjectivity can be a major contributor to obscuring true student ability. Clearly, many threats to the validity of résumé writing scores stem from the subjectivity of performance scores. It is thus of utmost importance to ensure that best practices are employed in prevention and identification of subjective errors. When sources of systematic error variance due to aspects other than student ability are identified, they can be taken into account and even eliminated (Linacre, 1989). In this thesis, the focus is on identifying the sources of résumé writing PA score error in hopes of improving the scoring process. The activity of systematic measurement error identification is part of gathering validity evidence for résumé rubric scores.

**Generalization**. In résumé writing PA situations, practitioners are interested in inferences about student résumé writing. Unfortunately, as described earlier, measurement error associated with PA measures is difficult to eliminate. By understanding the sources and amount of error, one has a better idea of the precision associated with PA scores. Generalization involves inferring the quantity of expected scores over all aspects of the measurement situation (Shavelson & Webb, 1991). Among other aspects, variability due to raters presents the concern of inter-rater reliability.

***Assumptions 3 and 4.*** *Raters produce consistent scores relative to each other* and *raters consistently agree on a program's score relative to the behavioral anchors on the Résumé Rubric*. To support generalizations to the universe score, researchers can conduct reliability studies or generalizability studies, both of which evaluate the consistency of scores across samples of observations (Kane, Crooks & Cohen, 1999). The universe of generalization is one of the subdomains of the target domain. Further, an individual's universe score is the expected score "over the universe of generalization" (Kane et al., 1999). Kane's terminology reflects the language used in Generalizability Theory (Brennan, 1997; Cronbach et al., 1972). Next, Classical Test Theory and Generalizability Theory in relation to the context of the Résumé Ruler are reviewed.

**Classical Test Theory reliability and precision**. In Classical Test Theory (CTT) a behavioral measure, X, is composed of the true underlying ability score, T, and error, e, which is considered to be due to random causes: $X = T + e$. In CTT, the error term is undifferentiated and considered random ($X = T + e_r$). The presence of random error ($e_r$) implies that residuals cancel out over all possible observations. In addition, it is assumed that true scores and error scores are uncorrelated, and that error scores from different measures are uncorrelated (Embretson & Hershberger, 1999). CTT provides reliability coefficients that allow the estimation of the degree to which the T component is present in a measurement. Reliability can be defined as a correlation between true scores and all possible observed scores that could be calculated from a person taking a test infinitesimally (Lord, Novick, & Birnbaum, 1968). Reliability may also be described in terms of the proportion of variance in true scores to the variance in the observed scores (Mellenbergh, 1996). Thus, variance in observed scores, the denominator of this

proportion, can consist of factors other than true score variance. From this perspective, reliability can be viewed as a complex term that is affected by many types of variance associated with a particular measurement situation; it is a sample-dependent estimate of measurement *precision* for a population. Because true scores fluctuate across samples, this "precision" coefficient reflects the consistency of scores within a particular sample and thus cannot be compared with measurement precision of another sample.

Several types of reliability exist in CTT because the error term is undifferentiated. Test –retest reliability provides information about the consistency of examinee test ranks over time. On the other hand, internal consistency measures the degree to which individual items in a test provide similar and consistent examinee scores. Further, parallel-forms reliability examines the rank-ordering of examinees by score across two alternative test forms. The error variance estimates vary depending on the reliability index of interest (Embretson & Hershberger, 1999). Because CTT provides only one definition of error, error due to different measurement elements of a research design is undifferentiated (e.g., items or occasions?).

An alternative measure of precision, the standard error of measurement (SEM) describes the standard deviation of errors of measurement associated with true score estimates derived from a sample of observed scores (Harvill 1991; Lord, Novick, & Birnbaum, 1968). SEM may be a more practical precision indicator than a reliability estimate because it refers to a specific type of variance, caused by the fluctuations of observed scores around the true score. Thus, it is not dependent on true score variability within a sample and can be conceptualized more accurately as score precision around a certain scale point. It is precision of measurement for a given subject (Mellenbergh,

1996). SEM can be useful when examining precision of scores associated with a scale under question. Because reliability is dependent upon the variability of scores in a particular sample, comparing reliability coefficients may not be as meaningful as comparing SEM across samples.

SEM can be reduced in context of relatively large test score variability in a given sample. In other words, measurement error is differentiated from the variability of scores in the sample (Mellenbergh, 1996). Conversely, in a sample of the same size with smaller score variability, reliability is reduced because scores are clustered more closely together, making rank ordering of scores less consistent. In this case, SEM can still be interpreted in context of the sample test score variability, and is therefore better suited to make comparisons of design consistency across groups. Given the distinction between precision indices, reducing rating measurement error associated with rubrics can be viewed both in terms of increasing reliability of the sample scores and decreasing SEM (Mellenbergh, 1996). Further discussion will reveal the connection of absolute and relative reliability to these concepts.

**Generalizability theory dependability.** Generalizability theory (G-theory) extends Classical Test Theory (CTT) in providing a mechanism for examining *dependability* of behavioral measurements (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). One of the main advantages of using G-Theory in establishing evidence of measurement soundness is that in this framework, the observed score can be partitioned into components other than the true test score and random error. Instead, it is possible to differentiate the classical error variability by partitioning it into constituents that represent variability due to the measurement situation, such as item difficulty, occasion

characteristics, and rater characteristics. Thus, G-Theory extends CTT in providing a mechanism for examining dependability of behavioral measurements (Cronbach, et al., 1972). A behavioral measurement such as a rubric score is considered a sample from a *universe of admissible observations*, which consists of all possible observations on an *object of measurement* (Webb, Shavelson, & Haertel, 2006). Types of variability due to measurement (e.g., items, occasions, raters) are called *facets* (Eckes, 2009; Shavelson & Webb, 1991). To estimate true score variance and error variance as well as possible, facets are identified during the measurement design identification stage. The identification of facets depends on the context of the measurement situation. For instance, when measuring student achievement on a writing exam, type of writing prompt may be a source of measurement error. Other measurement error sources of which the researcher is aware or suspects could be boredom levels, time of day, student's level of sleep, and so on. In the résumé writing review case, variability due to the difficulty of rubric criteria (i.e., elements) can be calculated and evaluated on its magnitude relative to variability due to the object of measurement. A large proportion of variance due to rubric element would indicate that in some elements it is more difficult to achieve a higher score than in others. On the other hand, a large proportion of variance due to raters indicates systematic rater harshness effects. It is also possible to estimate the effect of facet interactions (e.g., rater is harsh in some element ratings, but lenient in others).

In contrast to CTT, in the G-theory framework, the error term can be partitioned into systematic error and random error, $X = T + e_s + e_r$. The $e_s$ element represents facet variability that can be further partitioned depending on the number of facets involved in the research design. These systematic variances are called *variance components*, which

can be calculated and applied in determining the dependability of a measurement (Cronbach et al, 1972). In the résumé writing assessment design, variance components are associated with raters and elements facets. Systematic variance is also calculated for the object of measurement, *person.*

Similar to variables having values, facets are comprised of levels that can be defined as *random* or *fixed* (Shavelson & Webb, 1991). Random facets include levels that can be exchanged from the universe of generalization. Conceptually, a facet that is random indicates that the levels included in the analysis are an unbiased sample of levels that could be drawn from the universe of generalization (Cronbach et al., 1972). In the case of an objective test, an item facet is considered random if it is truly interchangeable with any other item measuring the same unidimensional trait. In the case of a PA rubric, a rubric element constitutes an "item." However, in PA cases, it is difficult to make the case that elements are interchangeable. For example, the quality of an objective statement is theoretically independent of the organization of a résumé, and thus a performance score on one element should not necessarily connote a performance score on another. Conversely, fixed facet levels represent the full theoretical scope of the facet and cannot be exchanged with any other level. A facet is fixed when the number of levels in its universe matches the observed number of levels (Shavelson & Webb, 1991). For example, imagine an alphabet test in which 26 items each represent a letter. In this case the measured levels exhaust the universe of generalization. The item facet, therefore, would be considered fixed. Fixed facets do not contribute systematic score variance to a fully-crossed design because they are held constant.

In the G-Theory framework, the object of measurement can be crossed with different facets. Crossing notation is such that if all essays in the sample are reviewed by all raters, $p \times r$. In a fully-crossed design, each level of every facet and the object of measurement are crossed. For example, all résumés can be crossed with all levels of the rater facet, which indicates that every rater provides rubric ratings for every résumé in the measurement situation. The object of measurement can also be nested in certain facets (Shavelson & Webb, 1991). The notation for essays being nested in raters is $p:r$. When the objects of measurement or facets are nested within the population of objects of measurement, it becomes more difficult to differentiate effects as they become confounded. For example, when sets of raters are nested within rater teams, the universe of admissible observations contains raters that are associated with only one rater team. Crossed designs are favored in generalizability studies, although nested designs are often used for convenience or for increasing sample size (Shavelson, Webb & Rowley, 1989). Increasing the sample size, in turn typically reduces estimated error variance and increases estimated generalizability (Shavelson, Webb & Rowley, 1989). Practically speaking, the optimal study design in some situations may well be the nested design.

Nested and crossed manipulations of the object of measurement, and random or fixed facets yield coefficients of dependability. Unlike CTT, G-Theory differentiates between *relative* and *absolute* reliability or dependability (Shavelson & Webb, 1991). Relative dependability refers to the consistency with which students can be ranked based on performance quality. For instance, résumé element scores can be ranked for each person across two or more raters; the consistency with which the raters rank the résumé quality of each person is relative to each résumé. This type of dependability is

represented by the G-coefficient. However, because in many cases considering absolute quality of résumé writing is more meaningful rather than simply comparing résumés across students, absolute dependability of a measure can be more relevant. Absolute dependability is consistency with which scores occur around a particular scale point. This dependability is represented by a Φ-coefficient. Thus, it is possible to determine consistency with which ratings from different raters occur around a specific quality point of résumé writing.

To determine the magnitude of Φ- and G-coefficients, one must first calculate both the relative and the absolute variances associated with the study design. Specifically, relative and absolute variances are calculated by adding the variance components from the G-study after adjusting for the levels associated with each D-study facet. Relative variance can be calculated by taking the sum of adjusted error variances that are directly related to the object of measurement, while absolute variance consists of all summed variances including those not due to or crossed with the object of measurement. Equations (1) and (2) respectively represent the relative and absolute variances of a fully-crossed design with a rater and element facets. It should be noted that the square root of $\hat{\sigma}^2_{\text{Abs}}$ (also $\hat{\sigma}^2_{(\Delta)}$) is the standard error of measurement (SEM).

$$\hat{\sigma}^2_{\text{Rel}} = \frac{\hat{\sigma}^2_{pr}}{n'_r} + \frac{\hat{\sigma}^2_{pi}}{n'_i} + \frac{\hat{\sigma}^2_{pri,e}}{n'_r n'_i} \quad [1]$$

$$\hat{\sigma}^2_{\text{Abs}} = \frac{\hat{\sigma}^2_{r}}{n'_r} + \frac{\hat{\sigma}^2_{i}}{n'_i} + \frac{\hat{\sigma}^2_{pr}}{n'_r} + \frac{\hat{\sigma}^2_{pi}}{n'_i} + \frac{\hat{\sigma}^2_{ri}}{n'_r n'_i} + \frac{\hat{\sigma}^2_{pri,e}}{n'_r n'_i} \quad [2]$$

*where $\hat{\sigma}^2_r$ is the rater facet variance component, $\hat{\sigma}^2_i$ is the element facet variance component, $\hat{\sigma}^2_{pr}$ is the person by rater interaction variance component, $\hat{\sigma}^2_{pi}$ is the person by element interaction variance component, $\hat{\sigma}^2_{ri}$ is the rater by*

*element interaction variance component, and* $\hat{\sigma}^2_{pri,e}$ *is the person by rater by element interaction confounded with random error variance and other unidentified sources of error*.

Using variances calculated with Equations 1 and 2, relative and absolute dependability coefficients for specific measurement designs can be estimated.

$$\hat{\rho}^2 = \frac{\hat{\sigma}^2_p}{(\hat{\sigma}^2_p + \hat{\sigma}^2_{REL})} \quad [3]$$

$$\phi = \frac{\hat{\sigma}^2_p}{(\hat{\sigma}^2_p + \hat{\sigma}^2_{ABS})} \quad [4]$$

Thus, if the measurement situation contained different occasions, several raters and different tasks, it is possible to obtain estimated variance components relative to each measurement element in the design. Instead of calculating different reliability coefficients as is in CTT, the variance components are used to compose an overall estimate of dependability of data, which takes into account the measurement variance accounted for in the design. Similar to conducting analyses of variance (ANOVA), it is possible to calculate the proportion of variance associated with each unwanted variance source to the overall variability in the data.

The ground for the generalization inference is the observed score. At this stage, Chapelle and colleagues (2010) recommend conducting generalizability and reliability studies as well as scaling and equating studies. Further, support for generalization to the expected performance can be made from standardizing task administration conditions. In the case of the Résumé Ruler, this is not applicable because there is only one task (i.e., one written résumé), and because students are free to complete the task at their own pace in a self-selected setting. For the Résumé Ruler, generalizability support can lead to the

conclusion that the observed scores reflect the expected résumé quality scores across raters.

**Adjusting for Rater Characteristics**

Given that several types of rater criteria effects typically present threats to the generalizability inference, it is important to attempt to identify and adjust any biases. Fortunately, several methodologies have been developed to address issues of differential severity of raters and the differential difficulty thresholds of rating criteria (e.g., Linacre & Wright, 2002; Raymond, Harik & Clauser, 2011). Further, equating methodologies for objective test equating can inform procedures in this study. The goal of test equating is to develop equivalent scores across different test forms (Kolen & Brennan, 2004; Lamprianou, 2008). Adjustments exist in Classical Test Theory (CTT) and Item Response Theory (IRT) frameworks; however the focus here was on CTT rater adjustments. Unlike in IRT, which estimates item-level parameters, in CTT the equating focus is on the total score (Kolen & Brennan, 2004). The data collection design used in this study resembles the nonequivalent groups anchor test (NEAT) design used in objective test equating (Petersen, Kolen & Hoover, 1993), in which different groups are administered different forms of a test, with a common set of items present in both forms (Kolen, 1988).

In typical objective test NEAT equating designs a set of common items is used to adjust for ability differences across different samples (Kolen & Brennan, 2004). In contrast, in the résumé assessment design a common set of résumés has been included across four different groups of raters in order to adjust for harshness/leniency of raters. Because different sets of raters rate different sets of résumés, without a common set of

résumés across all groups, it would be impossible to tell whether the set of raters is lenient or the set of résumés is of high quality (Raymond & Viswesvaran, 1993). In the rater bias situation, rubric scores would be inflated in that group, distorting the true pre to post score difference due to résumé writing appointments across all groups. Because the sample size of both résumés and raters is relatively small, adjustments in this study were demonstrated for illustrative purposes only. It is *not* recommended to make inferences from adjustments made with this small of a sample. However, the design of the study may be used on larger samples to make adjustments with high confidence about adjustment appropriateness.

If rater harshness is treated similarly to item difficulty, the methods described for objective test equating can be applied to performance situations (Raymond, Harik & Clauser, 2011). Specifically, because all rater teams are linked via anchor products, it is possible to scale all rater group résumé quality scores onto the metric of any one rater group. In this study, ratings across all four groups of raters were adjusted using the difference between the team mean anchor score and each group anchor means. This difference represents the relative harshness of each rater group. Then, each original score was adjusted using the constant for each group. Again, it is emphasized that the adjustment is used to demonstrate a technique that could be used with larger samples.

**Research Questions**

The Résumé Ruler is a performance assessment measurement tool designed specifically to align with programmatic objectives of a career office. Messick (1995) identified the structural integrity of the scoring framework as a critical aspect of validity for performance assessments. In order to make inferences from the Résumé Ruler scores, the measurement processes and structural quality of the instrument must be explored.

Given the need for additional validity evidence, this research serves two purposes. First, it can provide structure by which other rubric users can validate their own assessment processes. Second, the Résumé Rubric has the potential to be used by institutions outside of JMU, improving the overall field of career office assessment. To accomplish this second purpose, the ratings must be shown to be dependable when using a sample of raters appropriate for a variety of institutions outside of JMU. A more specific purpose to this last goal is to examine the generalizability of résumé quality scores. Gathering evidence for the generalizability inference allows the consideration of a more efficient programming designed to improve résumé writing.

This research provides additional validity evidence regarding the generalizability inferences with regard to rubric scores. The assumptions associated with this inference include that *raters provide consistent scores relative to each other and raters consistently agree on a program's score relative to the behavioral anchors on the Résumé Rubric*. This research also investigates the rank order of examinee abilities for the differential rater severity and differential scoring criteria difficulties. The following research questions were investigated in this study:

I. *How generalizable are résumé evaluation scores based on the current design?*

A. How generalizable are résumé ratings provided by professionals both in terms of relative and absolute decisions?

B. What are the variance component values associated with this design?

    a. What is the extent of variance due to rubric criteria in the résumé writing scores?

b. What is the extent of variance due to rater characteristics in the résumé writing scores?

C. What is the precision of rater's scores relative to the rubric criteria?

a. What is the standard error of measurement associated with the current design?

b. What is the typical range of standard error of measurement associated with the design in which there are three raters scoring the same 12 résumés?

*II. Based on pre-post scores on the résumé ruler rubric, how much do students' résumés improve?*

A. Are there statistically and practically significant value-added effects regarding each résumé element?

B. What are statistical and practical within-team changes between first and last résumés when scores are adjusted for team leniency/harshness?

CHAPTER THREE

# Method

## The Instrument

The Résumé Ruler is an instrument developed to measure student learning outcomes associated with participating in résumé writing appointments at James Madison University. Due to the fact that the instrument was described in the introduction section of this thesis, this section gives a brief overview. The Résumé Ruler consists of two parts: a checklist portion and an analytic portion. The checklist portion contains 25 components that should be present in most résumés: contact information (4 items), education (7 items), spelling/grammar (5 items), supplemental materials (3 items), and consistency (8 items). Consistency items include uses of font, punctuation, and format throughout the document. Each of the five checklist areas contained either two or three check boxes for each item – "No" (signifying an absence of a résumé element), "Yes" (signifying the presence of a résumé element) and "N/A" (signifying varying qualities explained within the checklist). An example N/A option includes, "N/A - More than 2 font styles is acceptable for targeted opportunity" in the Font checklist area. The checklist results can be used for program evaluation purposes because scores in this section result from less subjective decision-making than those in the analytic portion. Thus, advisors' attention to detail can be evaluated using this portion of the measurement tool. See Appendix B for the checklist.

The analytic portion of the Résumé Ruler contains six elements: objective statement, related experience, supporting/secondary experience, organization, headings, and appearance. The analytic components were rated on a 1 to 4 scale supplemented with

detailed behavioral anchors. Both portions of the measure were revised iteratively over the course of four years as résumé appointment practices changed.

In the past year, major changes were made to the Résumé Ruler. These changes need evaluation in order to continue giving helpful feedback to students and to make appropriate inferences based on Résumé Ruler scores for program evaluation purposes. Specifically, three analytic criteria were combined, resulting in one Supporting/Secondary section of the rubric. The three sections were Supporting/Secondary Experience, Awards, and Skills.

**Participants**

The participant raters in this study were 12 career and academic advisors employed in a career and academic planning office of a mid-sized southeastern university. Each – as a part of their positions - routinely met with students in one-on-one résumé improvement appointments. Nine of the career advisors were female and three were male. Advisors varied in résumé critiquing expertise from two years to ten years of experience.

Advisors were trained to use the current version of the résumé rubric during two separate training sessions. The training sessions included a general Résumé Ruler component overview as well as several practice résumé rating sessions. Advisors were asked to rate each of the practice résumés individually after which the advisors calibrated their responses. The résumés in the actual study belonged to two freshmen, five sophomores, seven juniors, and eight seniors.

**Procedures**

Résumés were collected by career and academic advisors in the last résumé review appointment with students. Résumés were sorted by an assessment graduate assistant into packets containing 12 résumés per advisor: six first draft (pre) résumés and six final draft (post) résumés. The participants constituted four rater groups with three raters in each group. The raters within each group rated an identical set of 12 résumés. Further, two résumés were rated by all raters across all four teams. Thus, overall 42 first and last résumés were rated: 40 within specific groups and two common to all groups. All groups rated a common pair of résumés in order to compare group leniency/harshness across groups. It is important to note that raters were not assigned last résumés corresponding to the same students' first résumés. In fact, to avoid bias that may arise when the same rater scores both the first and the last résumés for any one student, each rater team was assigned ten unique résumés (five first and five different last résumés). While avoiding bias due to exposure effects, the assessment coordinator can still acquire evidence of student growth from first to last résumé draft because a separate set of raters was assigned to rate last résumés versus first résumés. This design is represented in Figure 1. Rater team 1 scored ten pretest and posttest résumés that corresponded to ten posttest and pretest résumés rated by rater team 2. Similarly, rater team 3 scored ten résumés that corresponded to its respective counterpart résumés rated by rater team 4. Two résumés were rated by all teams. The résumé-rating session was conducted during a two-hour staff meeting during the summer academic semester.

| | Group1 (n=3) | Group2 (n=3) | Group3 (n=3) | Group4 (n=3) |
|---|---|---|---|---|
| Pre-test | 1 | 6 | 11 | 16 |
| | 2 | 7 | 12 | 17 |
| | 3 | 8 | 13 | 18 |
| | 4 | 9 | 14 | 19 |
| | 5 | 10 | 15 | 20 |
| | 21 | 21 | 21 | 21 |
| Post-test | 6 | 1 | 16 | 11 |
| | 7 | 2 | 17 | 12 |
| | 8 | 3 | 18 | 13 |
| | 9 | 4 | 19 | 14 |
| | 10 | 5 | 20 | 15 |
| | 22 | 22 | 22 | 22 |

*Figure 1.* The Résumé Ruler Study Design. The common set of résumés contains one first and one last résumé rated by each of the 12 raters.

Although nesting résumés in different rater teams increased the sample size of the observed résumés and allowed for unbiased ratings from first to last résumés, nesting also presents a problem. When résumés are nested within teams as is the case with the résumé review assessment situation, it may be challenging to detect rater harshness/leniency when inter-rater agreement within the team is high. For example, in the case of all raters in a particular rater group being equally harsh, a résumé would receive a rating lower than that reflecting the student's true ability to create an effective résumé. Anchor items are one way to overcome this limitation. By providing all raters with a set of the same résumés, it is possible to identify leniency/harshness issues within each rater team (Eckes, 2009). Thus, all four groups of raters received a common set of résumés composed of

one pretest résumé and one posttest résumé; however, longitudinal growth could not be calculated for this pair of résumés because they came from two different students.

Kolen and Brennan (2004) recommend anchor items on objective tests to constitute 20% of the total test length. Due to practical limitations, in this study the anchor products (i.e., résumés) constitute 17% of the total number of products; however this represents only two anchor résumés across all teams. The anchor products were included to be able to compare rater harshness across the four rater teams by adjusting résumé scores using the anchor mean for illustrative purposes. In this study, the anchor adjustments were not used to make inferences, but to demonstrate how individual students might be impacted by unadjusted versus adjusted scores.

**Analyses**

The ultimate goal of this study was to address validity issues regarding inferences made from assessment results that help a particular Student Affairs program make better decisions. Inferences about the substantive questions regarding the quality of an intervention are possible to the extent that the measurement tool is precise. Two coefficients were used to investigate the dependability of résumé ratings. However, more weight should be given to one of the two coefficients depending on the nature of the research question and the context within which dependability is interpreted. The G-coefficient ( $\hat{\rho}^2$ - coefficient) is interpreted when the research interest is in rank ordering individuals by score. In the case of the Résumé Ruler, if it were the case that the subject of interest was percentiles of student performance, or there was some other need to rank order student résumés by quality, the $\hat{\rho}^2$ - coefficient would be interpreted. The Phi-coefficient ($\phi$-coefficient) is interpreted when the research interest is in the absolute

score precision, or score consistency relative to the scoring scale. In the Résumé Appointment program the career office is interested in rating precision because the focus is on getting an accurate estimate of student learning due to résumé instruction. That is, there is more concern with ratings relative to the rubric, the Résumé Ruler. Because the point of collecting résumé ratings is to get a general idea of the program's strengths and weaknesses, and not to appraise students' relative performance, assessment quality is gauged and set by the rubric and thus scores relative to its criteria.

In addressing Research Question I (RQ I) - the generalizability of résumé scores - G-Theory was used to estimate variance components for sources of variance acknowledged to be important to the universe of admissible observations. The analyses used to evaluate rater dependability were conducted using the Generalized Analysis of Variance System (GENOVA; Crick & Brennan, 2001). Résumé Rubric element scores served as indicators for the construct 'quality of résumé writing', the object of measurement in the universe of observations. Using generalizability theory notation, the full design of the dataset is $\left[\left(p\times r\right):t\right]\times i$ which describes that each student résumé, *p* was crossed with raters, *r*, that were nested in teams, *t*, and each student résumé was rated on each of the elements, *i*. RQ IA- generalizability for relative and absolute decisions - was examined using $\hat{\rho}^2$ and $\phi$ coefficients, with an emphasis on $\phi$. Rubric criteria and rater characteristic variance components were compared to overall variability in the sample to address RQ IB – variance component contribution. Four D-studies were conducted to estimate variance components within each rater team. Each study fully crossed a set of 12 résumés' element scores with three raters. RQ ICa (precision associated with the current design) and RQ ICb (precision associated with a three rater

design) were addressed by calculating the absolute standard errors (SEMs) associated with the current design and SEMs associated with the design that used only three raters (i.e., one team), respectively.

It could be argued that the items facet in the Résumé Ruler experimental design can be treated either as fixed or as random. In the first line of reasoning, these particular measurement tool elements have been identified as the most important résumé elements by previous studies. The résumé elements in the analytic rubric of the measure may be the most essential and thus reflect the full universe of generalization. Thus, it can be argued that generalizing beyond the six elements is not theoretically warranted. Also, different elements represented in the Ruler may not be considered interchangeable given that certain elements are more "important" than others. For example, the quality of the "Related Experience" element may arguably be more important than the quality of the "Headings" element. The second line of reasoning contends that information represented by some of the rubric elements may become unimportant, and other facets not mentioned in the rubric may become more important to employers in the future. This second line of reasoning suggests that the items should be treated as random. Thus, the nested design analyses were conducted both ways to reflect the two alternative interpretations of the items.

Raters were represented as a random facet because raters should be interchangeable in the context of résumé rating. First, elements were considered as a fixed facet because the rubric was created to be representative of the content areas deemed most important to résumé content; thus, the construct of résumé writing quality was assumed to be represented in its entirety by these particular six elements. Elements

were expected to elicit different scores throughout the rubric because elements differ in difficulty. For instance, the mean rating for the Objective element across the 22 résumés was 1.65 (between "inadequate" and "below average"), whereas the mean rating for element Organization was 3.24 (between "above average" and "exceptionally executed"). This characteristic of the data is partially due to the problematic scoring rule for the Objective element: the behavioral anchor for a score of one reads "Objective is irrelevant to target OR is not included on résumé." Thus, if a student excluded an objective statement on either first or last résumé draft, she would automatically receive a "1." These differences in element difficulty are due to the theoretical nature of the construct and would introduce artificial error variance if elements were treated as a random facet (Orem, 2012). Nevertheless, elements were considered random in one of the analyses to demonstrate the impact of doing so and to generate discussion about generalizability.

To address RQ IIA - value added effects in each rubric element - a series of dependent t-tests were conducted using unadjusted and adjusted ratings. Overall and element value-added scores were evaluated for statistical and practical significance. Cohen's (1988) *d* criteria for small (.2), medium (.5), and large (.8) effects were used to judge the effect of the résumé writing appointments on student development overall and in the specific résumé element areas.

A focus of this study is the extent to which the adjustments for rater leniency/harshness would change résumé rubric scores. RQ IIB – practical value added with adjusted scores - was addressed by illustrating how individual scores were impacted by team-specific harshness and how the problem could be alleviated. Essentially, the anchor résumé scores can be used to center the data; that is, résumé scores can be

adjusted relative to the anchor résumé mean. Specifically, the difference between each rater team overall résumé anchor mean score and the anchor mean score across all teams can be used to adjust résumé ratings for team leniency/harshness. Again, because of the small sample of anchor résumés, the adjustments served as an example of the technique that could be applied when anchor products are more abundant.

CHAPTER FOUR

## Results

The results of this study are organized into major parts by research question (RQ): results related to the Résumé Ruler psychometric properties (i.e., RQs IA, IBa, IBb, IBc) and the results related to program evaluation (i.e., RQs IIA and IIB). The experimental design was $\left[\left(p\times r\right):t\right]\times i$ with three raters nested within four teams rating twelve résumés each, two of which are common to all four teams, and therefore to all twelve raters. Each rater evaluated six pre and six post résumés for different students across six elements of the rubric. In accord with the unraveling research questions nested within the generalization stage of the assessment score validation, the first analyses were partially nested design D-studies reflecting the number of facet levels used in the experimental design. Thus, RQ IA (relative and absolute dependability if résumé evaluation scores based on the current design) was answered by conducting two analyses to determine how dependability of this design differs when elements are regarded as fixed and random. Next, RQ IBa (rubric element variance component magnitude) was addressed by examining the partially nested design variance components associated with rubric elements. In addressing RQ IBb (rater variance component magnitude), variance components from the partially nested design and separate team variance components associated with raters were examined. RQ ICa (precision associated with the current design) and RQ ICb (precision associated with a three rater design) were addressed by calculating the absolute standard errors (SEMs) associated with the current design and SEMs associated with the design that used only three raters (i.e., one team), respectively. Finally, RQ IIA (value-added effects in each rubric element) and RQ IIB (adjusted value-

added effects) were addressed in the second major section by conducting dependent t-tests and performing a résumé score adjustment.

**Analyses Addressing Psychometric Properties of the Assessment Tool**

**Partially nested design analysis.** Because both the argument for fixed and for random element levels may be valid, $\hat{\rho}^2$ (G) and $\phi$ (Phi) coefficients were calculated for a design that treats the item facet as fixed and also for a design that treats the item facet as random. Each of the two D-studies modeled four levels of the random team facet, three levels of the random raters facet, and six levels of the fixed element facet. One should note that fixing the item facet will inevitably result in higher generalizability coefficients because the universe of generalization under consideration is smaller.

***Fixed element facet.*** Using generalizability notation, the full design of the dataset is $\left[\left(p\times r\right):t\right]\times i$ which indicates that each student résumé, *p*, was crossed with raters, *r*, that both were nested in teams, *t*, and each student résumé was rated on each of the fixed elements, *i*. Because résumés and raters were crossed, but nested in teams, RQIA (dependability of résumé ratings in terms of relative and absolute decisions) and RQIB (contributions of variance components associated with the partially nested design) were examined by conducting a partially nested analysis. Variance components for each facet and object of measurement (i.e., résumés) are presented in Table 1. Because GENOVA does not calculate dependability coefficients for nested models, $\hat{\rho}^2$ and $\phi$ coefficients were hand-calculated using formulae 5 and 7:

$$\hat{\rho}^2 = \frac{\hat{\sigma}^2_{p,pt}}{\hat{\sigma}^2_{p,pt} + \hat{\sigma}^2_{\delta}} \qquad \phi = \frac{\hat{\sigma}^2_{p,pt}}{\hat{\sigma}^2_{p,pt} + \hat{\sigma}^2_{\Delta}}$$

*where* [5,6] and *where* [7,8]

$$\hat{\sigma}^2_{\delta} = \frac{\hat{\sigma}^2_{pr,prt}}{n_r} \qquad \hat{\sigma}^2_{\Delta} = \frac{\hat{\sigma}^2_{t}}{n_t} + \frac{\hat{\sigma}^2_{r,rt}}{n_r} + \frac{\hat{\sigma}^2_{pr,prt}}{n_r}$$

Dependability coefficients were in line with acceptable ranges in applied research contexts, $\hat{\rho}^2$ = .89 and $\phi$ = .91. Although guidelines for acceptable values of $\phi$- and $\hat{\rho}^2$ -coefficients have not been established in the literature, it is justified to use the familiar Cronbach's α cutoffs. According to George and Mallery (2003), one can interpret Cronbach's α greater than .90 as "excellent", about .80 as "good", about .70 as "acceptable", about .60 as "questionable", about .50 as "poor," and anything less than .50 as "unacceptable." With these guidelines in mind, a $\hat{\rho}^2$ -coefficient value of .80 was interpreted as good relative dependability, whereas a $\phi$ -coefficient of the same value was considered representative of very good absolute dependability. Thus, an answer to RQ IA is that using the fixed elements design, the rubric ratings exhibited excellent relative and absolute dependability.

Next, focusing on conclusions RQ IBb (rater variance component magnitude), variance components associated with the partially nested design are described. The résumés-nested-within-teams variance component (i.e., $\hat{\sigma}^2_{p,pt}$, proxy for object of measurement) was .2685, and accounted for approximately 35.0 percent of total within-team variance in résumé ratings. In other words, approximately a third of the average total rating variability within teams was between résumés. In contrast, the raters-nested-within-teams variance component ($\hat{\sigma}^2_{r,rt}$ = .0199), which represented the amount of

variability due to differences in rater leniency/harshness within teams, accounted for only 2.6 percent of the average total rating variability within teams. The résumé x rater-nested-within-teams variance component ( $\hat{\sigma}^2_{pr,prt}$ =.0700) indicated that 9.1 percent of the average total rating variability within teams was due to differences in the relative rank order of résumés by raters. Thus in response to RQ IBb, systematic rater harshness and leniency was a relatively small issue within teams, on average.

The element x résumé-nested-within-teams variance component ( $\hat{\sigma}^2_{pi,pit}$ = .2115) represented 27.6 percent of average total rating variability within teams. This indicated that a large proportion of within-team variability was due to differential average relative element difficulty across résumés. In other words, within each team on average, raters as groups scored some résumés higher than average on some elements and other résumés higher on other elements. The rater x element-nested-within-teams facet ( $\hat{\sigma}^2_{ri,rit}$ =.0154), which illustrated the relative rater leniency/harshness on particular elements within teams, accounted for only 2.0 percent of the average total rating within-team variability. The final variance component related to within-team variance was the mixture of the variance attributable to the résumé x rater x element interaction within teams and random error still remaining in the model. This confounded variance component ( $\hat{\sigma}^2_{pri,prit}$ = .1818) accounted for 23.7 percent of average within team variability.

In addressing RQ IBa (rubric element variance component magnitude) the following variance components were relevant. The remaining variance components represented various facets of average variability between résumé scores in general. The variance component representing variability of average team scores ( $\hat{\sigma}^2_t$ = .0136) made

up 1.3 percent of total variability. In contrast, variability due to element difficulty ($\hat{\sigma}_i^2$ = .2674) comprised 25.4 percent of total variability. Finally, the element x team variance component ($\hat{\sigma}_{ti}^2$ = .0003) represented .3 percent of total variability, indicating little dependence of element difficulty on team membership. Thus, in response to RQ IBa, there were substantial differences between element scores across all raters and teams.

The estimate for score precision (standard error of measurement, SEM) was calculated by taking the square root of the absolute error variance component associated with the random elements design. The absolute SEM associated with this design was .18, indicating that raters within teams were on average about .18 points away from the universe score (RQ ICa, precision associated with the current design).

***Random element facet.*** When treating elements as random, $\hat{\rho}^2$ and $\varphi$ coefficients were hand-calculated using formulae 9 and 11. It should be noted that although the numerator in formula 9 ($\hat{\sigma}_{p,pt}^2$) is denoted by the same symbol as the numerator in formula 5, the résumé variance component in the fixed element design was equivalent to the sum of the random $\hat{\sigma}_{p,pt}^2$ and $\frac{\hat{\sigma}_{pi,pit}^2}{n_i}$ variance components. This is due to the fact that the element variance (crossed with persons and nested within teams) was not accounted for by any facet variability in the fixed design, and thus remained as "person variance".

$$\hat{\rho}^2 = \frac{\hat{\sigma}_{p,pt}^2}{\hat{\sigma}_{p,pt}^2 + \hat{\sigma}_{\delta}^2}$$

*where*

$$\hat{\sigma}_{\delta}^2 = \frac{\hat{\sigma}_{pr,prt}^2}{n_r} + \frac{\hat{\sigma}_{pi,pit}^2}{n_i} + \frac{\hat{\sigma}_{pri,prit,e}^2}{n_r n_i} \qquad [9,10]$$

and

$$\phi = \frac{\hat{\sigma}^2_{p,pt}}{\hat{\sigma}^2_{p,pt} + \hat{\sigma}^2_{\Delta}}$$

*where* [11,12]

$$\sigma^2_{\Delta} = \frac{\hat{\sigma}^2_{t}}{n_t} + \frac{\hat{\sigma}^2_{r,rt}}{n_r} + \frac{\hat{\sigma}^2_{i}}{n_i} + \frac{\hat{\sigma}^2_{ti}}{n_i} + \frac{\hat{\sigma}^2_{pi,pit}}{n_r} + \frac{\hat{\sigma}^2_{ri,rit}}{n_i} + \frac{\hat{\sigma}^2_{pr,prt}}{n_i n_r} + \frac{\hat{\sigma}^2_{pri,prit}}{n_i n_r}$$

Treating the element facet as random reduced the estimated dependability of Résumé Ruler ratings however, dependability was still acceptable. In comparison with the fixed design, relative dependability decreased to $\hat{\rho}^2$ = .80, whereas absolute dependability decreased to $\phi$ = .67. The increased interval between the relative and absolute dependability coefficients simply indicated that when considering a universe of interchangeable résumé rubric elements, element difficulty made a greater impact on the object of measurement to total variability ratio. And, as evident from the fixed element analysis, reducing the number of rubric elements to which we generalize had a positive effect on the dependability. This makes sense because in the latter design, element difficulty became part of the error. Thus, an alternative answer to RQ IA is that when using the random elements design, the rubric ratings exhibited good relative dependability and acceptable absolute dependability.

The estimate for score precision was calculated by taking the square root of the absolute standard error variance associated with the random elements design. The absolute SEM associated with this design was .34, indicating that raters within teams were on average about .34 points away from the universe score (RQ ICa, precision associated with the current design).

Table 1 displays variance component contributions to total and average within-team variability. Using the random element approach yielded similar variance components. The remaining unexplained variability decreased by about 5 percent; relative résumé x item variance within teams also decreased by approximately 6 percent. Because total estimated variability has decreased, but each variance component changed little, the decreases in relative importance of explained variability is arbitrary.

The percent variability in Table 1 was calculated differently for nested sources of variation (p:t, r:t, pr:t, pi:t, ri:t, and pri:t,e) and free sources of variation (t, i, ti) because variability within teams has a different meaning than variability between teams. Whereas the *variability within teams* for any one of the nested sources of variation is the average variability across four teams, *variability between teams* represents the actual variance between teams and items. For example, the magnitude of the "team" (t) variability denotes differences between teams' average résumé scores. That is, variance is examined holistically. On the other hand, the "raters within teams" (r:t) source of variation represents not how *all* raters' average résumé scores vary across the entire design, but the average of how raters vary within each team. Thus, percent total variance for each of the nested sources of variation (p:t, r:t, pr:t, pi:t, ri:t, and pri:t,e) was calculated separately using the denominator that is the sum of the nested variance components. The percent total variance for the free sources of variation (t, i, ti) was calculated out of the total observed score variability. In the fixed elements design, total "nested" variability was .7671, whereas total variability was 1.0512. In the random elements design, total "nested" variability was .6992, whereas total variability was .9827.

Table 1

*2011-2012 Résumé Ruler Ratings Using the Partially Nested Design: Contribution of Each Facet to Score Variance*

| | | Items Fixed | | | Items Random | | |
|---|---|---|---|---|---|---|---|
| Source of variation | Notation | Variance Component | SE | % Total Variance* | Variance Component | SE | % Total Variance* |
| Résumés within teams (p:t) | $\hat{\sigma}^2_{p,pt}$ | 0.2685 | 0.061 | 35.0 | 0.2333 | 0.061 | 33.4 |
| Raters within teams (r:t) | $\hat{\sigma}^2_{r,rt}$ | 0.0199 | 0.012 | 2.6 | 0.0174 | 0.012 | 2.5 |
| Team (t) | $\hat{\sigma}^2_{t}$ | 0.0136 | 0.029 | 1.3 | 0.0130 | 0.029 | 1.3 |
| Items (i) | $\hat{\sigma}^2_{i}$ | 0.2674 | - | 25.4 | 0.2674 | 0.147 | 27.2 |
| ti | $\hat{\sigma}^2_{ti}$ | 0.0031 | 0.011 | 0.3 | 0.0031 | 0.011 | 0.3 |
| pr:t | $\hat{\sigma}^2_{pr,prt}$ | 0.0700 | 0.010 | 9.1 | 0.0398 | 0.011 | 5.7 |
| pi:t | $\hat{\sigma}^2_{pi,pit}$ | 0.2115 | 0.026 | 27.6 | 0.2115 | 0.026 | 30.3 |
| ri:t | $\hat{\sigma}^2_{ri,rit}$ | 0.0154 | 0.007 | 2.0 | 0.0154 | 0.007 | 2.2 |
| pri:t,e | $\hat{\sigma}^2_{pri,prit}$ | 0.1818 | 0.012 | 23.7 | 0.1818 | 0.012 | 26.0 |

*Note.* *Variance components' % Total Variance for facets that were nested within the teams facet were calculated using total within team variance only. % Total Variance for non-nested components was calculated from the summed total of all variance components. The standard error (se) for the items facet (when fixed) was not calculated because the effect is not generalized to other samples, which precludes consideration of sampling distributions.

**Disaggregated analyses of rater dependability by team.** Although dependability coefficients were at least acceptable in the full D-studies, the variance components contributing to total variability within teams may differ depending on the team. Further, it is informative to learn about the typical variance component contributions when three raters score 12 artifacts. Four fully-crossed D-studies modeling three levels of the random rater facet (*r*) and six levels of the fixed element facet (*i*) were conducted to examine the within-team variability more closely. Dependability coefficients and absolute standard errors of measurement (SEMs) were calculated to describe dependability of résumé ratings. Team SEMs summarize each team's rating

precision. Team SEM can be interpreted as the overall team ratings' average distance from the team's universe scores. In other words, the extent to which the raters are close to the team average score is an indication of how close ratings are to each other on average within each team. Absolute SEM represents the precision of a team's résumé set ratings relative to the rubric behavioral anchors. The absolute SEMs were calculated for each team by taking the square root of the absolute variance component from a D- study design using three raters and six fixed elements. Table 2 contains the summary of variance component contributions in each of the four rater teams.

The following is the type of information available for scrutiny about each of the rater teams. On an ordinal scale of 1 to 4, Team 1 assigned an average résumé score of 2.62 (i.e., between "below average" and "above average") to the combined set of six initial (pre) and six final (post) résumé drafts. Team 1 members were consistent in rank-ordering résumés by quality with each other ( $\hat{\rho}^2$ = .87), and had good consistency relative to the rubric scale ( $\phi$ = .84). On average, raters were 0.29 points away from the résumé rating universe score. About 84 percent of rating variability was due to differences in résumé quality and only 3.5 percent of rating variability was due to systematic rater harshness/leniency. This indicates that most of the differences in résumé quality scores were due to actual differences in résumé quality and not construct-irrelevant effects such as rater leniency or confusion due to rubric element ambiguity.

Table 2 describes variance component contributions for all four teams. Team members were fairly consistent with each other ( $\hat{\rho}^2$ range .77 - .89), and had mostly adequate consistency relative to the scale ( $\phi$ range .52 - .84). Three raters within each team varied in distance from the résumé quality universe score (RQICb, precision

associated with a three rater design). Standard error of measurement associated with the overall résumé score ranged from .2492 to .3425. Notably, Team 2 had low total variability in résumé quality ratings, which may have contributed to the low absolute dependability coefficient. For Team 2, only approximately 52 percent of score variability in the rubric scores was due to résumé quality, whereas 33.9 percent were due to systematic rater leniency/harshness (RQIBb, rater variance component magnitude). That is, almost half of the differences in résumé quality was due to rater leniency/harshness effects and misinterpretation of rubric elements or other unidentified sources of error. In such cases detecting true résumé quality scores is difficult, and scores are considered less dependable than ratings with a higher percent of variability attributed to résumé quality differences. It should be noted that on average, this group's average résumé score was still similar to that of Team 1. Team 3 had the highest overall mean across all résumés.

Table 2

*Variance Component Contributions within Each Rater Team in Four Fully-Crossed, Fixed Element Design Studies*

| | $\hat{\sigma}_p^2$ | $\hat{\sigma}_p^2$ % | $\hat{\sigma}_r^2$ | $\hat{\sigma}_r^2$ % | $\hat{\sigma}_{pr,e}^2$ | $\hat{\sigma}_{pr,e}^2$ % | $\hat{\rho}^2$ | $\phi$ | Total Variance |
|---|---|---|---|---|---|---|---|---|---|
| Team 1 | 0.4450 | 84.3 | 0.0182 | 3.5 | 0.0648 | 12.3 | 0.87 | 0.84 | 0.5694 |
| Team 2 | 0.1271 | 52.0 | 0.0828 | 33.9 | 0.0345 | 14.1 | 0.79 | 0.52 | 0.2444 |
| Team 3 | 0.2087 | 68.7 | 0.0336 | 11.1 | 0.0617 | 20.3 | 0.77 | 0.69 | 0.3040 |
| Team 4 | 0.2005 | 76.4 | 0.0368 | 14.0 | 0.0253 | 9.6 | 0.89 | 0.76 | 0.2626 |

*Note.* Standard Error of Measurement (SEM) calculated using $\sigma_\Delta^2$ associated with each team's rating variability was 0.2880 for Team 1, 0.3425 for Team 2, 0.3087 for Team 3, and 0.2492 for Team 4.

One important point to take into account is that dependability estimates depend on not only the relative variability due to a particular facet, but also on the amount of overall variability present in the design. In the team-specific D- studies elements were fixed, causing the only difference between relative variance and absolute variance associated

with each D-study to be the rater leniency/harshness variance component. Thus, the differences between the $\hat{\rho}^2$'s and the $\phi$'s was due to the rater leniency component ($\hat{\sigma}_r^2$). For example, total variability in Team 1 (.5694) doubled that of Team 4 (.2626); Team 1's rater variance component impact on absolute error (3.5%) was a quarter of Team 4's (14%). The effect on absolute dependability was even more drastic when small total variability in rater scores is compounded with a higher rater variance component. This is illustrated by the example of Team 2, which had low total variability and a higher rater variance component. These examples serve to better understand the dependability coefficients.

**Element impact**. To gain a better understanding of the role elements play in Résumé Rubric rating dependability, 24 team-by-team D-studies were conducted investigating one element at a time. These studies also provided precision around element score information. In investigating rubric element difficulty, rubric elements were ranked according to the element mean magnitude (see Table 3 for element difficulty rank ordering). Whereas Team 1, Team 2 and Team 4 raters agreed on the average ranking of elements by difficulty, Team 3, characterized by a high $\hat{\sigma}^2_{pr,e}$ component (i.e., résumé x rater interaction confounded with random error), ranked element difficulties differently. The Objective element (rank order = 5) and Related Experience (rank order = 2) elements were ranked consistently by difficulty across all four teams. It should be noted that notwithstanding similar overall average rubric scores (2.62, 2.64, and 2.46 for Team 1, 2, and 4, respectively), the dependability of rubric scores from Team 2 was relatively lower in terms of consistency relative to the rubric anchors.

One may note that the large overall mean and the lack of accord in element difficulty rank ordering could be due to sampling error. Sampling error could affect: 1) better quality résumés being assigned to Team 3 by chance (hence, the larger team mean), a different profile of strengths and weaknesses within the résumé set assigned to Team 3 (hence, the different element rank order of element means), or 3) both.

Table 3

*Rank-Order of Rubric Elements by Mean Score*

| | Team 1 | | Team 2 | | Team 3 | | Team 4 | |
|---|---|---|---|---|---|---|---|---|
| Element | *M* | Rank | *M* | Rank | *M* | Rank | *M* | Rank |
| Objective | 2.34 | 5 | 2.36 | 5 | 2.78 | 5 | 2.33 | 5 |
| Related Experience | 2.82 | 2 | 3.11 | 2 | 3.19 | 2 | 2.53 | 2 |
| Supporting/Secondary Experience | 2.31 | 6 | 2.08 | 6 | 2.96 | 4 | 2.33 | 6 |
| Résumé Organization | 2.82 | 3 | 2.58 | 3 | 3.24 | 1 | 2.53 | 3 |
| Headings | 2.38 | 4 | 2.54 | 4 | 2.61 | 6 | 2.39 | 4 |
| Appearance | 3.03 | 1 | 3.15 | 1 | 3.00 | 3 | 2.65 | 1 |
| Total | 2.62 | | 2.64 | | 2.96 | | 2.46 | |

*Note.* Rating "1" (Section is inadequate and requires an overhaul OR is not included); "2" (Section is below average and requires a good deal of improvement before submitting); "3" (Section is above average and needs minimal improvement); "4" (Section is well done and is exceptionally executed). The rank order was determined by giving the highest rank to the element with the highest mean score (descending order).

**Rating precision related to elements**. Standard errors of measurement (SEM) due to elements summarize raters' precision around each element across all résumés. The element SEM can be interpreted as the precision of the score based on a single element. Each element's absolute SEMs were analyzed for each rater team (see Table 4). Absolute SEM represents the precision of element ratings relative to the rubric behavioral anchors. In each of the 24 D-studies the rater facet was the only modeled source of interpretable systematic error; each study described the rater variance components associated with each element. In addressing RQ ICb, (precision associated with a three rater design)

absolute SEMs were calculated for each element by taking the square root of the absolute variance component from a D- study design using three raters.

Teams 2 and 3 appeared to have similar standard errors across the six elements; Résumé Organization had the smallest standard error ($SE_{Team2}$ = .33, $SE_{Team3}$ = .28), whereas Objective ($SE_{Team2}$ = .67, $SE_{Team3}$ = .63) and Headings ($SE_{Team2}$ = .73, $SE_{Team3}$ = .57) had the largest standard error for these two teams. Teams 1 and 4 had the smallest standard errors associated with Appearance ($SE_{Team1}$ = .25, $SE_{Team4}$ = .24) and Résumé Organization ($SE_{Team1}$ = .26, $SE_{Team4}$ = .25). For these two teams, the largest standard errors were associated with Supporting/Secondary Experience ($SE_{Team1}$ = .60, $SE_{Team4}$ = .55).

Table 4 provides the absolute standard errors for all four teams on all six elements. Overall, the general trends were that Résumé Organization (SE range .25 - .33), Appearance (SE range .24 - .36), and Related Experience (SE range .32 - .43) had the smallest standard errors across all four teams. In contrast, the largest standard errors were associated with Supporting/Secondary Experience (SE range .55 - .60), Objective (SE range .47 - .67) and Headings (SE range .52 - .73).

Table 4

*Rank-Order of Rubric Elements by Element Absolute Standard Errors of the Mean*

| | Team 1 | | Team 2 | | Team 3 | | Team 4 | |
|---|---|---|---|---|---|---|---|---|
| Element | *SE* | Rank | *SE* | Rank | *SE* | Rank | *SE* | Rank |
| Objective | 0.5912 | 5 | 0.6667 | 5 | 0.6310 | 6 | 0.4665 | 4 |
| Related Experience | 0.3568 | 3 | 0.4342 | 3 | 0.3600 | 2 | 0.3263 | 3 |
| Supporting/Secondary Experience | 0.5990 | 6 | 0.5693 | 4 | 0.5465 | 4 | 0.5528 | 6 |
| Résumé Organization | 0.2631 | 2 | 0.3333 | 1 | 0.2817 | 1 | 0.2546 | 2 |
| Headings | 0.5204 | 4 | 0.7312 | 6 | 0.5693 | 5 | 0.5159 | 5 |
| Appearance | 0.2546 | 1 | 0.3402 | 2 | 0.3624 | 3 | 0.2379 | 1 |

*Note.* The rank order is determined by ascending order of SEMs.

## Value-Added Effects

The confidence with which assessment practitioners can generalize performance assessment results depends on how consistent raters were relative to each other and relative to the rubric in scoring student résumés. Thus, it is important to examine the amount of rubric score variability due to rater leniency/harshness, element criteria ambiguity, and any interactions between the rater and element facets and résumés. In relation to group-level decision making, this research suggests the résumé ratings are appropriately dependable. Therefore, I embarked on the second phase of the analysis.

**Statistical and practical significance of unadjusted scores.** To address RQ IIA (statistical and practical significance of value-added effects regarding each résumé element), a series of dependent samples t-tests were conducted to determine whether student résumé elements differed from pre-résumé consultation to post-résumé consultation. Cohen's *d*'s were calculated to provide an indicator of practical effect size. Specifically, it indicates the magnitude of differences between pre- and post-scores in standard deviation units. See standard benchmarks for the traditional Cohen's *d* (Cohen, 1988) in Table 5.

Table 5

*Benchmarks for Traditional Cohen's d*

| Value | Effect |
|---|---|
| 0.2 | Small |
| 0.5 | Moderate |
| 0.8 | Large |
| > 1.0 | Very Large |

The overall résumé score averages differed significantly, $t(59) = 5.94$, $p < .001$. In addressing RQ IIA (value-added effects in each rubric element), both statistical and

practical significance is relevant. To this end, with the exception of Objective, each element displayed statistically significant and practical gains from first to last drafts of the résumé. Further, with the exception of Objective, each element was associated with small to very large practical gains. Table 6 displays the overall and element-specific results of the paired comparison. The trustworthiness of these results depends at least in part on the dependability of each of the element ratings. Therefore, it is important to remember that ratings associated with Supporting/Secondary Experience, Objective, and Headings were associated with diminished precision (see Table 7 for detailed precision information in the context of gain scores on each element). It is possible that rater imprecision associated with these elements attenuated gains within these areas. It is also possible that students are truly improving less in Supporting/Secondary Experience and Objective areas and improving greatly in creating appropriate Headings due to instruction.

Table 6

*Means, Standard Deviations, Statistical and Practical Significance of Differences between Pre-test and Post-test Résumé Rubric Ratings*

| Rubric Component | *df* | Pre-Mean | *Pre-SD* | Post-Mean | *Post-SD* | *Diff* | *t* | *d* |
|---|---|---|---|---|---|---|---|---|
| Content | | | | | | | | |
| Objective Statement | 59 | 1.52 | 0.89 | 1.55 | 0.91 | 0.03 | 0.22 | 0.03 |
| Related Experience | 59 | 2.65 | 0.82 | 3.17 | 0.73 | 0.52 | 4.84** | 0.62 |
| Supporting/Secondary Experience[a] | 59 | 2.68 | 0.87 | 3.07 | 0.82 | 0.39 | 2.74* | 0.36 |
| Format | | | | | | | | |
| Résumé Organization | 59 | 2.57 | 0.95 | 3.21 | 0.82 | 0.64 | 4.75** | 0.61 |
| Headings | 59 | 2.25 | 0.74 | 3.23 | 0.76 | 0.98 | 5.27** | 1.04 |
| Appearance[b] | | | | | | | | |
| Appearance | 59 | 2.57 | 0.85 | 3.18 | 0.72 | 0.61 | 4.34** | 0.57 |
| Total | 59 | 2.37 | 0.63 | 2.90 | 0.46 | 0.53 | 5.78** | 0.85 |

*Note.* [a] least reliable; [b] most reliable; *$p < .01$; **$p < .001$.

**Anchor score-based adjustments.** Overall, students appeared to improve their résumé writing score with the final draft. In the Résumé Appointments program students' first and last résumé scores were aggregated across teams to provide an overall rubric gain score. This "impact score" describes the average student résumé quality improvement exhibited across all Résumé Ruler elements. However, if the program was interested in providing students with individualized feedback based on assessment scores, how accurate would these scores be? Individual teams' average harshness/leniency had the potential to obscure true résumé quality gains because different teams with different harshness/leniency levels rated the same students' résumés. Thus, it was possible that a

lenient team provided overly high post-test scores for the same students that received attenuated pre-test scores from a harsh team. In this situation, gain scores for that particular group of résumés would appear spuriously impressive. If résumés were not shared across teams, it would be difficult to determine whether teams were harsh/lenient, or résumés allotted to those particular teams were of low/high quality.

Because anchor résumés were scored by each of the rater teams, it was possible to adjust team-specific rubric score gains using the anchor scores. Specifically, each original average résumé score was linearly transformed using the following two simple steps: 1) each team's average anchor score was subtracted from the overall average anchor score, and 2) the difference was added to each original average résumé rating, taking into account team membership. For example, if the overall anchor score was 2.5 and Team 1's average anchor score was 2.75, each résumé score rated by Team 1 raters would be adjusted *down* for team leniency by .25 points. The resulting adjusted average résumé ratings would represent résumé quality adjusted for each team's harshness/leniency relative to the other teams. As a reminder, these adjustments were conducted for purely illustrative reasons due to the low anchor sample size. Table 7 illustrates the anchor score means and overall adjustment values for each team.

Table 7

*Team-Specific Anchor Averages and Adjustments*

| | Team 1 | | Team 2 | | Team 3 | | Team 4 | |
|---|---|---|---|---|---|---|---|---|
| | Mean | ADJ | Mean | ADJ | Mean | ADJ | Mean | ADJ |
| Overall | 2.75 | -0.0035 | 2.88 | -0.1285 | 2.97 | -0.2257 | 2.39 | 0.3576 |
| First Résumé | 2.14 | 0.0139 | 2.31 | -0.1528 | 2.31 | -0.1528 | 1.86 | 0.2917 |
| Last Résumé | 3.36 | -0.0208 | 3.44 | -0.1042 | 3.64 | -0.2986 | 2.92 | 0.4236 |

*Note.* Overall, first résumé and last résumé adjustments were calculated by subtracting each team's appropriate anchor mean from the grand anchor mean (2.75). The team-specific adjustments were applied to each average résumé score.

Again, assuming that a substantial number of anchors was included, another possible adjustment would be take into account harshness/leniency that was specific to first résumés and final résumés separately. Theoretically, systematic harshness/leniency could look very different relative to final résumés than relative to first résumés due to rubric criteria floor/celling effects. Further, the latter differences could depend on team membership. To account for these differential harshness/leniency effects, assessment practitioners could 1) subtract each team's average first/last résumé anchor score from the overall average first/last résumé anchor score, and 2) add this difference to each original résumé rating taking into account team membership and pre/posttest status. Continuing with the previous example, let the overall anchor score equal 2.5, Team 1's first résumé anchor score equal 2.5 and last résumé anchor score equal 3.0. Each first résumé score provided by Team 1 raters would be left intact, whereas final résumé scores would be adjusted *down* by half of a rubric scale point. This substantial difference in Team 1 scores would then contribute to deflating the gain score for résumés 6 through 10. The advantage of this technique is that it allows adjustments that take into account not only team harshness/leniency, but differential harshness/leniency across first and last résumés. Table 7 illustrates first résumé and last résumé adjustment values for each team. Table 8

illustrates the adjustment summary for average team anchor scores and average first and last team anchor scores.

Table 8

*Rubric Element Gain Scores with and without Pre-test and Post-test Anchor Mean Adjustment*

| | Résumés 1-5 Gain | Résumés 6-10 Gain | Résumés 11-15 Gain | Résumés 16-20 Gain |
|---|---|---|---|---|
| Original | 0.48 | 0.51 | 0.92 | 0.02 |
| Adjusted[a] | 0.62 | 0.38 | 0.36 | 0.55 |
| Adjusted[b] | 0.58 | 0.42 | 0.22 | 0.69 |

*Note.* [a] Overall team anchor adjustment; [b] First and last anchor résumé adjustment.

# CHAPTER FIVE

## Discussion

This study investigated psychometric properties of a performance assessment tool used for program evaluation purposes. Specifically, a résumé rubric developed by a career office on a four-year university campus was used to guide students in the process of improving their technical writing skills. In addition, the rubric was used to later score students' pre-consultation and post-consultation résumés in order to evaluate the efficacy of student instruction during one-on-one résumé review appointments. In order to make appropriate inferences regarding assessment results, assessment practitioners must take into account rating unreliability associated with performance assessments. Although many methods are available, generalizability studies are among the most appropriate options for performance assessment score evaluations. In this thesis, several generalizability analyses were used to investigate the relative and absolute dependability associated with résumé rubric scores.

Similar to the results organization, this discussion was organized into two major sections: technical considerations (i.e., RQ I, How generalizable are résumé evaluation scores based on the current design?) and applied considerations (i.e., RQII, Based on pre-post scores on the résumé ruler rubric, how much do students' résumés improve?). With reference to the original need for validating assessment tools, technical considerations include steps that can be taken to improve the dependability of evaluation scores. Further, given that precision was adequate and even of desirable extent (depending on which approach was taken with the elements facet – fixed or random), next steps in Kane's recommended validation process are discussed. With reference to the stakeholder

needs, applied considerations include the presentation of results in such a way that program changes based on the results are possible. In other words, how can we shed light on the fact that program effectiveness differs by rubric element? Given the results, can student instruction, rater training, and the measurement tool be modified to yield a clearer picture of student gains due to Résumé instruction?

**Technical Considerations**

In this study the $\phi$- coefficient associated with the partially nested design, fixed element fact, appeared to be adequate for program evaluation purposes (i.e., overall $\hat{\rho}^2$ = .91 and $\phi$ = .89). This reflects good dependability when three raters in four teams rate randomly selected and assigned résumés. However, because there was confounding due to the nested nature of the rubric data, it was useful to examine each rater team separately. When separate team D- studies were conducted to examine sets of 12 résumés rated by three raters, some instability in the dependability coefficients was revealed. Team-specific $\phi$-coefficients ranged from .52 (poor) to .84 (good), whereas absolute SEMs ranged from .25 to .34, indicating that within teams specific résumé scores may be more biased than the overall dependability estimate initially suggested. Thus, it was imperative to investigate the typical measurement error sources within teams.

**Limitations of the nested design.** Although the $\phi$-coefficient is the more appropriate dependability estimate in the context of the inferences made from student rubric scores, it is limited in that it is interpreted in the context of a nested design. Thus, because overall more raters and résumés are considered in its calculation, dependability appears high. However, facet variability was averaged over four teams, and thus was

confounded with teams to a certain extent. For example, if the rater variance component was low in Team 1 but high in Team 2, the variance component in a design nesting raters within these two teams would look more like an average of the two.

Three possible solutions could address the issue with using the nested design. First, instead of having four separate rater teams, one could conduct a fully crossed study with twelve raters scoring twelve résumés. In generalizability studies it is favorable to use a fully-crossed design (Brennan, 1992), which would allow a more precise estimate of the typical rater and rubric element effects on rubric scores The drawback to fully-crossed designs is that the object of interest must be crossed with every facet. This would present a resource issue in that many raters would be rating only a limited number of résumés. This in turn increases the chance of résumé sampling error playing a part in the assessment results. That is, if the smaller sample of résumés happened by chance to include mostly high-quality artifacts, then rater and element variance components would not be representative of the full possible range of résumé quality. Second, one could have four rater teams rating twelve different sets of résumés. This approach, however, prevents making inferences regarding relative team harshness or résumé quality across teams. It is problematic that raters within any one team with the highest résumé scores could happen to be the most lenient team, or could have by chance received résumés that were better in quality. A third option, then, would be to include anchor résumés that would be rated by all teams. With a large enough sample of anchor résumés, this mechanism would allow comparisons of résumé quality across all teams. The anchor design would be especially useful in situations where fairness to individual students was

of importance or when training and rubric development are either not feasible or do not produce noticeable increases in rating consistency.

**Fairness.** If résumé scores were to be shared with students, one would need to take into account fairness regarding the leniency and harshness of raters. Imagine a simple scenario where two résumés of equal quality were rated by two different raters, one harsh and one lenient. It is likely that the résumé evaluated by the harsh evaluator would be lower, which may be considered unfair. Fortunately, there are some measurement techniques that can adjust for such systematic error. One of the simpler techniques that allow for a fairer evaluation of individual student skill is mean equating. Mean equating provides a mechanism for equating scores provided that the distributions of scores from different [teams] contrast only in means and standard deviations (Muraki, Hombo, Lee, 2000). For example, anchor résumés could be used to adjust team scores by the difference between the overall anchor score and the team anchor score. This technique was applied to Résumé Ruler scores for illustrative purposes only.

To illustrate how anchor résumés could inform this assessment study, consider rater Team 3 and Team 4. Although only two anchor résumés were included in this study, the following example was included to demonstrate the usefulness and relative simplicity of adjusting scores based on anchors. To make such adjustments, one would want many more anchor résumés. Recall that in this study, teams differed substantially in mean résumé scores. Earlier, the issue of confounding raters with teams was described. Confounding prevents teasing apart whether some teams received better quality sets of résumés or some teams contained more lenient raters. So, how would one go about using

anchor résumés in investigating how biased individual résumé scores are due to team-specific levels of harshness/leniency?

The design was such that Team 3 rated first draft résumés (11-15) and Team 4 rated the final drafts of the same five résumés (see Figure 1 for a review of the design). The same pattern was set with the remaining two teams. Thus, no one team rated both initial and final versions of the same résumé. Teams were also blind to whether they were rating initial or final drafts. Element analyses results revealed that some elements exhibited increases in low posttest scores like "1" (inadequate) and "2" (below average). For example, Team 4 rated résumés 16-20 at pretest (M = 2.85, "Average") whereas Team 3 rated the same five résumés at posttest (M = 2.82, "Average").

If résumé quality was examined at team level, no overall changes in résumé quality would be detected. However, combining this information with anchor résumé means can determine whether résumés 16-20 actually improved from the initial to final drafts. Team 4 pretest anchor scores were on average .29 points lower than the average anchor pretest, and Team 4 posttest scores were on average .42 points lower than the average anchor posttest résumé score! On the other hand, Team 3 was on average .15 points higher than the average pretest anchor score, and .30 points higher than the average posttest anchor score. Thus, it is possible that in these teams student résumé writing skill development was washed out by the combined Team 4's relative harshness and Team 3's relative leniency. Thus, in a high stakes situation, students associated with résumés 16-20 would appear to have made no gains in résumé writing quality. Assuming that team-level leniency/harshness patterns hold, résumés 11-15 would exhibit the opposite effect as the pretest and posttest résumés were reversed between the teams.

Thus, résumés 11-15 would exhibit extraordinary value-added effects because now the harsh team would be scoring pre-test résumés and the lenient team – posttest résumés. Regardless of the effect, the true gain scores are masked entirely by rater-specific effects.

Because including anchor résumés in all teams allows for between-group comparisons regarding rater leniency/harshness, it is possible to adjust scores based on anchors. A more careful examination of such patterns, then, could potentially untangle the confounding between teams and résumés. To resolve this individual fairness issue, a simple solution would be to adjust the appropriate résumé groups for the amount by which each team differs from average pretest and posttest anchor scores. This could be done by hand; however sophisticated statistical packages exist that will automatically adjust for rater effects using a complicated set of algorithms that utilize the same basic principles demonstrated in this example. Using these techniques requires relatively large samples of performances and raters (which are clearly lacking in this small program assessment study).

In modern measurement theory, a family of more sophisticated and fine-tuned techniques is grouped under multifaceted Rasch modeling (MFRM). Within MFRM, Multi-faceted Rating Scale Model (MFRSM) (Linacre, 1989) studies the probability that a specific examinee will be rated with a specific rating scale point by a specific rater on a specific element. MFRMs depict the additive contribution of each element of the measurement context to the log odds of observing one rating scale category versus the next lower rating scale category. In this process, parameters that represent the object of measurement, typically a person, and facets of measurement context such as raters and items are used (Linacre, 1989). The probability may depend on the typical four

parameters: the examinee's proficiency, rater harshness, item difficulty, and threshold between two adjacent rating scale levels (Wolfe, 2009). This technique has been used in many large-scale performance assessments to adjust for the common measurement effects on individual performance scores (see Eckes, 2009).

Although examples of measurement techniques designed to address rater biases that training cannot correct are mostly limited to large scale assessment, (e.g., Eckes, 2009; Engelhard, 1994; Lumley & McNamara, 1995; Myford & Wolfe, 2002; Yi et al., 1997) rater bias needs to be identified and adjusted in smaller-scale situations. Researchers recommend the use of statistical models to detect and correct for rater and task effects in an effort to reduce systematic error in estimating performance assessment scores (Raymond, Harik, & Clauser, 2011). Since the current study is small in sample size and individual student performance is not of concern to the Résumé Ruler assessment program, such complex adjustments are not justified or necessary. However, in order to diagnose areas for improvement in the program and the assessment itself, assessment coordinators should understand element-specific score trustworthiness.

**Facet considerations.** Other than the overall dependability of Résumé Rubric scores, a major consideration of this study was how the element and rater facets contributed to overall score variability. This was useful because examining raters in isolation revealed the most direct information regarding rater training. Likewise, examining elements separately contributed the most direct information about improving the measurement tool.

Particularly, conducting separate D- studies by team made it possible to determine the typical variance components for rater effects when the elements were considered

finite. Modeling three raters, the rater variance component contribution to overall variability ranged from as small as 3.5 percent (Team 1) to 33.9 percent (Team 2). To get an idea of these rater effect magnitudes in relation to the rubric, a typical range of average rater scores can be constructed by taking the square root of each the rater variance component and multiplying it by four. If the square root of the rater variance component can be interpreted as a standard deviation around a rubric scale point, then two standard deviations above and below a résumé score will encompass about 95% of average ratings. Thus, the typical range of rater means across all résumés and elements for teams could be anywhere from .54 points on the four-point rubric scale (i.e., about half a rubric scale point) to as high as 1.15 points. The four different rater teams also differed in rating precision. Standard errors of measurement for each team ranged from .2493 (Team 4) to .3425 (Team 2).

An examination of standard errors of measurement associated with the element-specific D-studies using three raters uncovered several areas recommended for improvement. The absolute standard errors associated with elements-specific analyses spoke to the precision with which raters scored each rubric element on average (in each team). Across the four teams, raters agreed best on Résumé Organization and Appearance, but were less precise in rating Headings, Objectives, and Supporting/Secondary Experience.

**Strategies to improve reliability.** Using precision information supplied by absolute SEMs it is possible to develop strategies to improve reliability associated with the Supporting/Secondary Experience, Headings, and Objective rubric elements.

Imprecision issues can stem from ambiguities associated with the measurement tool, or training.

***Training improvement.*** More time spent explicating the elements in need of clarification could be a solution. During training, raters could individually review sample résumés and then discuss them with a partner, focusing on Supporting/Secondary Experiences, and Headings, in particular. Any differences in interpretation should be discussed and appropriate adjustments to the scoring rubric should be negotiated. Although this negotiation process can be time consuming, it can also greatly enhance reliability (Yancey, 1999). The Objectives element should also be addressed with a special focus through training; however there is reason to believe that this element should be improved by correcting its scoring procedure. If the Objective remains as a rated element on the rubric, then training could be improved by providing extra instruction on how to develop students' job or career objectives.

Further, greater extent of control can be introduced in any rating training situation by employing opportunities for continuous calibration of raters. Adjudication may eradicate issues with rater leniency, rater by element interaction, halo effect, and overall consistency relative to rubric elements. Rater training is an important practice and must be an element of the measurement process to the extent that it is feasible. Without training, raters may be mystified by rubric elements containing phrases that require additional definition (Campbell, 1999).

Taking these suggestions into account, the current training could be improved by organizing raters into groups, and supplying them with student products of various quality. The current organization of training involves the entire staff discussing their

ratings, which could create ambiguity and confusion. It is more efficient to ask staff to follow a three stage rating procedure where the first step is for raters to score a set of student products individually. Then, raters could discuss their decisions with a partner, and change original ratings due to this conversation. Finally, the full staff as a group may discuss the ratings in detail with a facilitator whereby any drastic discrepancies could be noted. The facilitator should refer frequently to the rubric criteria to resolve incongruity between staff members. This process may not only improve the consistency of rubric ratings, but also may shed light on rubric areas in need for clarification. In training, special consideration should be given to areas such as Related Experience, since raters appear to be less consistent within this element.

A positive byproduct of this study is a well-studied set of résumés that could be used as *range finders*. The New York State Education Department defines range finding as "reviewing and scoring students' field tested constructed responses to select anchor papers for the test scoring guides" (NYC Office of State Assessment, 2013). Similarly, the résumés scored for this program evaluation can be used in training. For example, during the training practice, raters can compare their scores to the average "known" scores observed in this study. Because the résumé scores ranged from approximately 1 to approximately 4 (full range), résumés that averaged at a specific overall criteria level (e.g., 3, "above average") can be used to represent clear-cut résumé quality. For example, an anchor résumé rated by all twelve raters around the score "3" could be an excellent example of what raters should look for in an above average student résumé.

***Rubric improvement.*** Well-designed scoring rubrics respond to the concern of intra-rater reliability by establishing a description of the scoring criteria in advance

(Moskal & Leydens, 2000). Intra-rater reliability refers to the extent to which each rater's scoring process changes over time. In these cases, inconsistencies in the scoring process result from influences that are internal to the rater rather than true differences in student performances. Moskal and Leydens (2000) recommend that establishing clear scoring criteria will allow raters to refer to constant scoring rules, and emphasizing the importance of frequently revisiting the rubric criteria can maintain consistency within each rater (i.e., decrease the rater x element interaction variance components). Indeed, beyond training, a major typical source of error in performance assessments is the rubric. That is, rater inconsistencies can be a product of insufficient clarity in rubric criteria. Thus, a brief discussion of possible measurement improvements follows.

From a measurement tool improvement standpoint, what are some steps that can be taken to improve the Résumé Rubric? One way to address this question involves differentially weighing various rubric elements. Although the justification for including each of the considered rubric elements is supported according to the literature (e.g., Tsai et al., 2011; Hornsby & Smith, 1995), some of the elements may be of less importance than others. Even though weighing less agreed-upon element scores will not improve rating precision for those elements, it can decrease an element rating's relative impact on the overall résumé score. For example, given that Supporting/Secondary Experience may not be as informative to an employer as Related Experience, and that raters agree less within this element, giving this element half the weight of the Related Experience element may be justified.

The relative lack of precision with respect to the Supporting/Secondary Experience element could be due to a specific set of factors. First, this element has been

constructed from three different criteria: awards, honors, and skills. Thus, this element is a complex criterion with several potentially separate elements combined into one. It may be difficult for raters to simultaneously score students' performance in all three areas at the same time. For example, a student's supporting or secondary skills section may be lacking, but she has clearly and concisely presented her skills and awards. To improve precision, the program coordinators should consider separating the rubric element into three elements of lesser weight. In doing so, raters will be able to consider each résumé component without also having to consider other parts. Second, because the three criteria were recently combined, a larger part of the training process may need to be devoted to this element. If the Supporting/Secondary Experience element is retained in its current form, a greater part of training should be spent on distinguishing between the rubric rules for each criterion.

Element "Objectives" was also associated with low rater precision. In this case, the lack of rater consistency was artificial in that two separate definitions existed for the same score (1). The first definition was that the "Objective is irrelevant to target", whereas the second definition was that the objective is not included on résumé. Thus, the probability of getting a score of "1" was greater than it would be if only one definition was used. Indeed, preliminary item analyses revealed that 41/60 (68.3 percent) total Objective pre-test résumés and 40/60 (66.7 percent) total Objective post-test résumés received a score of "1." These distributions are in stark contrast with those of other elements. A score of "1" was assigned to 6.7 percent pretest and 1.7 percent posttest résumés on Related Experiences; to 6.7 percent pretest and posttest résumés on Secondary/Supporting Experience; to 8.3 percent pretest and 25 percent posttest résumés

on Organization; to 20 percent of posttest résumés on Headings; and 6.7 percent pretest and 18.3 percent posttest résumés on Appearance. The prevalence of low scores on the Objective element simultaneously invalidates comparisons between elements and also using this element's score to calculate the average résumé score. Thus, program coordinators should consider excluding the Objective element from the calculation of the average résumé score if the objective truly is missing, and if an objective is irrelevant to the quality of that résumé.

**Applied Considerations**

After establishing that résumé rubric score dependability was adequate for aggregated pre-post comparisons, the résumé quality gain scores from pre-test to post-test were examined. Student résumés appeared to improve substantially from pretest to posttest. Indeed, there were statistically and practically significant value-added effects in each résumé element with the exception of Objectives. Other than in Objectives, post-test ratings across elements all exceeded "3, section is above average and needs minimal improvement." "Above average" ratings indicate that the Résumé Appointment program is doing an excellent job in student instruction overall. However, not all résumé areas exhibited homogenous gains. With precision information in hand, it is possible to make more accurate inferences regarding the differential gains across rubric elements. Moreover, it is important to establish the nature of the inferences we can make given the methodology used in this assessment design. What is the extent to which assessment coordinators can make causal inferences regarding value added effects given the design used? Further, other than "noise" created by measurement error associated with raters

and elements, what other variables can help explain the variability in post-test résumé scores?

Although generalizability analyses do not remove measurement error from calculations of construct score gains, one can strategically apply precision information to increase confidence in statements about intervention quality. In other words, rater precision around element scores can inform the inferences about résumé instruction effects and strengthen consultation. For example, without measurement information, smaller element-specific gains could indicate that there is room for improvement in instruction. However, before developing ways to improve instruction in those areas, revisiting element-score precision can increase confidence in or raise flags about observed gains. For instance, although statistically significant, gains associated with Résumé Ruler Supporting/Secondary Experience element were among the smallest ($d$ = .36). It may be useful to recall that Supporting/Secondary Experience was also associated with consistently relatively low rating precision across all teams (Table 4). Thus, true gain scores may have been obscured by the noise created around the score by measurement error. On the other end of the gain score spectrum, Headings exhibited a very large practical effect ($d$ = 1.04). However, the Headings element was also associated with some imprecision in ratings. Unlike Supporting/Secondary Experience, the gain in scores given relatively high imprecision is impressive. As a result of combining measurement and gain information, an assessment coordinator might recommend to postpone making final judgments regarding changes to the Supporting/Secondary Experience student instruction. The reasoning behind this is that it may be more worthwhile to focus on rubric development and rater training where raters

exhibited imprecision relative to the rubric scale. Of course, advisor training should still communicate detailed information regarding the standard instruction in all areas covered by the rubric.

In areas where rater precision was adequate (e.g., Appearance, Résumé Organization elements), a next step may be to attempt to increase gain scores by improving advisor training. In relation to this, it may be useful for program coordinators to set a desirable score that can be used to compare assessment results across assessment cycles. The practice of setting a desired score allows program coordinators to diagnose areas recommended for improvement. A desired score could be a standard such as a particular amount of score gain produced by instruction, or an average score that students must exhibit at post-test (e.g., students are expected to exhibit at least a level 3 average post-test rating, 'Section is above average and needs minimal improvement'). The choice of the desired score type depends on the nature of the program. Because of floor and ceiling effects of ordinal scale ratings, it may make more sense to establish a standard like a desired average post-test résumé rating. In this case, all students are expected to exit the program with a particular level of skill. Alternatively, gain score standards leave the possibility of adequate gains, but substandard performance. In addition, students who started with high scores could only make small gains. Since the Résumé Rubric has been under development, the assessment program coordinators have not yet developed a standard for desired gains. However, a standard of "3" appears to be reasonable given that with the exclusion of Objectives, all areas exceeded this score at post-test. Again, keeping in mind measurement precision information regarding the Objectives element,

one can make an argument that the lack of observed improvement in this area is an artifact of measurement error, or noise.

The Résumé Appointment is a program shown to be effective in improving an important employability skill. However, only 40 students were represented in the available pool of résumés in one academic year. It would appear that the career office should apply greater marketing efforts to reach more students with this successful program. On the other hand, the Résumé Appointment is only one part of a program portfolio containing other interventions such as résumé workshops and employer résumé reviews. At this time those programs are not associated with assessment. Given the observed impact of the résumé appointments, one could make an argument for initiating assessments corresponding with other programming. Further, recent literature supports the view that it is no longer adequate for college students to master the knowledge and skill within a particular content area – increasingly, gaining practical skills that will enhance graduates' chances of employment have become of great prominence (Fallows & Steven, 2000). Indeed, résumés are way for students to express employability skills beyond discipline-specific knowledge (Miner, 2000).

Although the effects demonstrated by the Résumé Appointment assessment program are impressive, they must be viewed from a larger methodological perspective. In other words, the evidence provided for program effectiveness must be viewed also from the perspectives of threats to internal validity. Threats to internal validity may be especially relevant in assessment practice, as the inferences made from results are encouraged to be put to practice (i.e., closing the loop). When an assessment coordinator measures the effect of a résumé program on résumé writing student learning outcomes,

she may make an inference that increased résumé writing proficiency is due to the intervention. Proposed programmatic changes based on this inference are appropriate to the extent that the causal relationship between the intervention and change is true student learning can be attributed to the intervention. Thus, it is paramount to have a high level of confidence that the positive observed effects are indeed due to the implemented programming and a lack of effects is due to an ineffective program. In other words, internal validity evidence is very relevant in the context of assessment practice.

Research design in assessment typically occurs in a quasi-experimental context. A quasi –experimental design is essentially an experimental design that does not meet requirements necessary for control of extraneous variables (Shadish, Cook, & Campbell, 2002). There is often little opportunity to randomly assign students to, for instance, different variations of the same program to judge which yields more useful learning and developmental outcomes. In Student Affairs assessment contexts, we typically make inferences about the effects of programming using measurements from a group of students that took part in our program. Instead of assuming that the measured student outcome changes are the direct result of the implemented interventions, it is best to make an argument for making such inference.

It is likely that the mechanisms behind student outcome change include extraneous variables that may bias program effectiveness inferences. Natural maturation, variation in program implementation across instructors, natural disasters, and economic shifts can confound assessment results. As a consequence, inferences about the real added value to student experience may be biased. To counteract biases in research, experimental and quasi-experimental researchers utilize “control” techniques. Common

mechanisms that foster control in experimentation include the use of random assignment and control groups (Pearl, 2000). Random assignment allows researchers to make claims that assigned groups are equivalent on relevant characteristics. For example, including a control group in an assessment design can allow the program coordinator to compare gain scores from students randomly assigned to an intervention to gain scores of students that did not receive an intervention. With such an arrangement, it is possible to rule out threats to internal validity such as natural maturation (e.g., students naturally become more professional in their technical writing).

In assessment practice, many other threats remain (e.g., students receive technical writing training in other courses); however by utilizing best methodological practices that chip away at the internal validity threats, it is possible to at least increase the confidence in the causal inferences validity. For a more detailed treatment of threats to internal validity, consult Shadish, et al. (2002).

In the case of the Résumé Appointments program, it appears that the program is making strides in getting students the information and guidance they need to organize professional information; however several threats to internal validity may be present. Of interest would be information regarding college student progression through liberal education regarding ability to effectively present personal professional information. For example, do students improve résumé writing skills without instruction? This question could be answered by examining samples of student work at several points in students' education. Demonstrating that résumé instruction results in increased résumé writing quality improvement rates represents powerful evidence for Résumé Appointment

effectiveness. Other, more accessible information can be obtained to achieve equivalent confidence gain in causal inference.

An organized way of thinking about information geared at increasing confidence in causal inference is to examine intervention fidelity. Intervention fidelity has been defined as having two components: 1) the extent to which those giving the intervention/program give the intended service and 2) the extent to which the students accept this service fully (Hulleman & Cordray, 2009). The essential purpose of conducting intervention fidelity checks is to counteract not meeting program objectives due to discrepancies between planned program implementation and actual implementation. A fidelity checklist is yet another rubric that may ask independent raters to investigate whether a program was implemented for the planned duration, whether the implementers adhered to addressing specific features for a program, and the extent of implementation quality. For smaller Student Affairs programs, other information is relevant to fidelity. In the context of Résumé Appointments, one may ask questions such as "How much time did students spend on résumés between consultations?" Alternatively, it would be interesting to find the average extent of improvement for students who attend more than two appointment sessions. That is, as exposure to the treatment increases, do students gain more? On the other hand, if students do not specifically ask for help on particular résumé characteristics, do students still receive this relevant information? Knowing that each résumé element gets attention during sessions can inform the origins of differential gains across Résumé Rubric elements. Results revealed that large gains were observed for Headings, but only moderate gains for Appearance. One hypothesis for this differential impact is differential time on these

topics. Specifically, perhaps consultants spend more time on headings than on appearance. One could only test such a hypothesis using some type of fidelity check.

In general, fidelity checks are a worthwhile endeavor that greatly facilitates program decisions regarding the question of most import – does the program work? Program coordinators should always seek additional information regarding the consistency between the hypothetical program and its reality. Although dealing specifically with threats to causal inferences validity, implementation fidelity studies contribute to validity of inferences regarding assessment results as does the main topic of this thesis – measurement dependability.

**Future Validity Considerations**

In this thesis, validity of assessment inferences was introduced through the lens of Kane's (1992) validity argument approach. The first three stages of Kane's argument-based approach to test score validation were addressed – domain description, evaluation, and generalization. The extent to which assumptions were met dictates whether or not the next stages of Résumé Rubric validation should be initiated. Reliability must be established before more advanced stages of validity can be examined.

First, the domain of résumé quality was defined. The assumption underlying domain description is that the *observations used to draw inferences about skill in the résumé writing domain involve performances on tasks from the domain of interest.* To evaluate this assumption, essential résumé elements were described and mapped to the existing résumé rubric used in this study. According to the literature review the résumé writing artifacts should reveal relevant knowledge, skills, and abilities in situations representative of those in the target domain of professional résumé writing.

Next, the assumption underlying the evaluation stage was investigated: *The criteria used to score the performance are appropriate and have been applied as intended*. This assumption was evaluated using generalizability analyses, which examined the precision associated with each rubric element and rater effects. The scoring criteria differed in appropriateness (e.g., Objective criteria lacked rules applied to other rubric elements), but as a whole appeared to be applied as intended as evidenced by generally good absolute measurement precision. Whereas three of the rubric components displayed adequate standard errors of measurement, others were associated with high discrepancies between raters. When twelve raters were divided into four teams that each rated twelve résumés, the general, element-specific and person-specific rater leniency/harshness was minimal (*r:t* effect: 2.6 percent; *pr:t* effect: 9.1 percent; *ri:t* effect: 2.0 percent). However, within individual teams rater-specific effects accounted for 3.5 percent to 33.9 percent of total variability. Thus, there is variability in the appropriateness of the scoring criteria application. Standard errors associated with elements informed how element characteristics affected ratings and several areas for improvement were uncovered (e.g., Secondary/Supporting Experience). However, it is expected that minor changes to the rubric and specific emphases on the more ambiguous rubric elements during training will eradicate the more major issues with rating precision.

According to this study, overall career advisors were adequately consistent relative to each other and also relative to the rating scale of the rubric. Thus, assumptions 3 and 4 (*Raters produce consistent scores relative to each other* and *raters consistently agree on a program's score relative to the behavioral anchors on the Résumé Rubric*) associated with the generalization inference were satisfied. First, advisors produced

consistent scores relative to each other (overall $\hat{\rho}^2$ = .91; team-specific $\hat{\rho}^2$'s .77 - .89, between good and excellent) and mostly agreed on résumé scores relative to the behavioral anchors on the JMU Résumé Rubric ($\phi$ = .89; team specific $\phi$'s .52 - .84). Although several minor improvements may be introduced into the rubric, career advisors were found to have sufficient understanding of most elements' anchor meanings.

The next step in the argument-based approach is the validation of the explanation inference. During this stage, one goes through the process of providing evidence that expected scores are attributed to a particular construct. That is, are the expected Résumé Rubric scores actually attributed to the construct of résumé writing ability? Among assumptions related to this inference is that résumé content knowledge and writing skill required to create a professional résumé will vary across experienced and novice technical writers. This assumption can be tested by comparing Résumé Rubric scores between the office student educators and first-year college students. Alternatively, performance on the Résumé Ruler should relate to performance on other measures of résumé writing quality (e.g., employment status six months after graduation). Further, the internal structure on the Résumé Ruler ratings is consistent with theory. To test the latter assumption, it is possible to utilize structural equation modeling (SEM). SEM corrects estimated relationships among latent variables for the biasing effects of measurement error due to items/elements. However, generalizability theory can be combined with structural equation techniques to account for errors due to raters (see DeShon, 1999).

The extrapolation inference is allowed when the researcher can show that assessment scores can account for the level of proficiency that would be perceived in a

professional setting. That is, the rubric scores must be shown to represent quality in the target domain of a professional résumé review. The assumption that can relate to this inference is that résumé appointment résumé performance is related to other contexts where résumé quality is important. To this end, one could investigate improvements to either the rubric or training, and manipulate conditions to assess improvement effects in each. For instance, after making the necessary changes to the rubric, one may investigate whether the psychometric properties have improved the ability to detect gains across particular elements. A between-groups experimental design may be used to evaluate whether the dependability and generalizability of the measure improved.

In conclusion, there are many possible avenues for increasing confidence in the validity of Résumé Rubric assessment scores. Validation of an inference is a continual process and is one of the most important endeavors in assessment practice. Measurement problems could lead to an undesirable scenario of stakeholders completely missing the true programmatic effects and making inappropriate resource decisions. With monitored measurement quality, one can disentangle true gains in student learning from noise and make the most appropriate programmatic decisions. By continuing to address measurement and research design with each assessment cycle, we ensure that implemented training is relevant and stakeholders are making decisions using data that represents what is most important. With some development the Résumé Rubric can serve as a standardized tool that other institutions with similar student career development goals can utilize in their own practice. Finally, the methodology described in this study may be useful to other institutions that have accepted assessment as one of the most important tools to building a successful body of college graduates.

# Appendix A

## The Résumé Ruler Rubric Analytic Criteria

| | ***RÉSUMÉ RULER* Analytic Criteria** | | | | | **ID:** |
|---|---|---|---|---|---|---|
| | General Scoring Guidelines | **Section is well done and is exceptionally executed.** | **Section is above average and needs minimal improvement.** | **Section is below average and requires a good deal of improvement before submitting.** | **Section is inadequate and requires an overhaul OR is not included.** | **SCORES** |
| **CONTENT** | **Objective** | Entire objective is specific and clearly relates to target. Strengths/skills are communicated using concise language. **4** | Entire objective is not specific and/or does not clearly relate to target. Strengths/skills are not included or are not communicated using concise language. **3** | Objective is a narrative or generic and target is vague or omitted. Strengths/skills are not included or are not communicated using concise language. **2** | Objective is irrelevant to target OR is not included on résumé. **1** | |
| **CONTENT** | **Related Experience** (Any section including relevant experience regardless of heading name or order in which it falls on the résumé) | All experiences are relevant and clearly relate to target. Experience descriptions use action verb phrases, are well-defined, and include specific skill-focused /concrete examples. Titles, locations and time frames are included for each experience. **4** | Some but not all experiences are relevant and/or do not clearly relate to target. Experience descriptions might not use action verb phrases, are less defined and/or do not include specific skill-focused /concrete examples. Titles, locations, and time frames may or may not be included for each experience. **3** | Most if not all experiences are not relevant and/or do not clearly relate to target. Experience descriptions do not use action verb phrases, are not defined and/or do not use specific skill-focused /concrete examples. Titles, locations, and time frames may or may not be included for each experience. **2** | Related experience is not included on the résumé. **1** | |
| **CONTENT** | **Supporting/ Secondary Experience** (Any experiences that are not considered "Related Experience" including "Awards", "Honors" and/or "Skills") | S/S experiences showcase writer as well-rounded, are well-defined, and add value to the overall résumé. Titles, locations and/or time frames are included, if applicable. Skills include level of expertise (i.e., familiar, proficient, etc.) when appropriate. **4** | S/S experiences might not showcase writer as well-rounded, not all are well-defined, and some do not add value to the overall résumé. Titles, locations and/or time frames are included, if applicable. Skills include level of expertise (i.e., familiar, proficient, etc.) when appropriate. **3** | S/S experiences fail to showcase writer as well-rounded, all are not well-defined, and do not add value to overall résumé. Titles, locations and/or time frames are omitted when applicable. Skills do not include level of expertise (i.e., familiar, proficient, etc.) when appropriate. **2** | All S/S experiences are irrelevant OR not included on résumé. **1** | |

| | | | | | | |
|---|---|---|---|---|---|---|
| FORMAT | **Résumé Organization** | Content is arranged so the most relevant experiences to target are first or above the less relevant experiences. **4** | Content is arranged so some of the most relevant experiences to target were placed below less relevant experiences. **3** | Content is arranged so it is difficult to determine which experiences are more or less relevant to target. **2** | Content is not organized, (i.e. it is written in paragraphs, outlined as a list, displayed in columns, etc.) **1** | |
| | **Headings** | Headings are specific and/or relate to target. **4** | Some headings are broad, vague and/or do not relate to target. **3** | All headings are broad, vague and do not relate to target. **2** | There are no headings on the résumé. **1** | |
| APPEARANCE | **Appearance** | Layout of content, highlighting features, and font used directs the reader's attention and gives the résumé a neat and professional appearance. **4** | Layout of content, highlighting features, and font used misdirects the reader's attention and gives the résumé a less professional appearance. (i.e. page is overcrowded, highlighting features are used inconsistently, and/or font distracts from résumé) **3** | Layout of content, highlighting features, and font used distracts the reader's attention and gives the résumé an unprofessional appearance. (i.e. page displays large areas of "white space", highlighting features may or may not be used, and/or font distracts from résumé) **2** | There does not appear to be any type of logical layout, use of highlighting features, and/or professional font. **1** | |
| | | OFFICE USE ONLY: | | **Total Analytic Score:** | | |

Appendix B
The Résumé Ruler Rubric Checklist Criteria

**RÉSUMÉ RULER CHECKLIST**

**Contact Information**

| | Yes | No |
|---|---|---|
| Name | ☐ | ☐ |
| Address where student can easily be contacted | ☐ | ☐ |
| Professional email address | ☐ | ☐ |
| Phone number where student can easily be contacted | ☐ | ☐ |

**Education**

| | Yes | No | N/A |
|---|---|---|---|
| Institution | ☐ | ☐ | |
| Institution Location (city, state) | ☐ | ☐ | |
| Degree | ☐ | ☐ | |
| Date of graduation | ☐ | ☐ | |
| Major/minor/concentration | ☐ | ☐ | |

**Supplemental Materials**

| | Yes | No | N/A |
|---|---|---|---|
| Reference Page | ☐ | ☐ | ☐ |

Yes - Brought separate reference page

No - References are listed on résumé

N/A - Neither of above

| | Yes | No | N/A |
|---|---|---|---|
| Portfolio, writing/media sample, etc. | ☐ | ☐ | ☐ |

Yes - Referenced

No - Not referenced but a necessary selection criteria (Associated with opportunities in fine arts, media, theatre, design, or writing)

N/A - Not referenced and not necessary selection criteria

| | Yes | No | N/A |
|---|---|---|---|
| Reference to website | ☐ | ☐ | ☐ |

Yes - Referenced

No - Not referenced but a necessary selection criteria (Associated with opportunities in computing, media, geographic science, mathematics)

N/A - Not referenced and not necessary selection criteria

**Consistency (Uses in Consistent Manner throughout the Entire Document)**

Note: If more than 1 institution is included, mark "Yes" only if all "Education" criteria is met for each institution. (Include past institution(s) when a degree has been earned and/or experience(s) at the past institution is referenced on the résumé)

GPA

Yes- GPA listed

No - GPA not listed but a necessary selection criteria

N/A -GPA not listed and is not necessary selection criteria

Certifications/Licensure - Month/Year

Yes - Any is listed. Remind to add month/year if not included

No - Not listed but relevant

N/A - Not listed but not relevant

**Spelling/Grammar - Quick Skim by the Reviewer**

| | Yes | No |
|---|---|---|
| **Error free** document (correct spelling and grammar) | ☐ | ☐ |
| Free of narrative descriptions (as used in cover letters) | ☐ | ☐ |
| Free of pronouns (other than in the objective) | ☐ | ☐ |
| Variety of action verbs/skills | ☐ | ☐ |

**Margins**

| | Yes | No | N/A |
|---|---|---|---|
| Font - One font style is recommended, two is acceptable | ☐ | ☐ | ☐ |
| Yes - No more than 2 font styles | | | |
| No - More than 2 font styles, negatively impacting appearance | | | |
| N/A - More than 2 font styles is acceptable for targeted opportunity | | | |
| Punctuation | ☐ | ☐ | |
| Highlighting features | ☐ | ☐ | |
| Use of bolding, italics, underlining, caps consistent within sections | | | |
| Placement and format of titles | ☐ | ☐ | |
| Consistent alignment of titles and order of position and organization | | | |
| Locations (city, state or JMU designation) | ☐ | ☐ | |
| City and state or JMU required for employment and related experience | | | |
| Date designation and placement | ☐ | ☐ | |
| E.g., use of 06/08 vs. June 2008 is consistent within sections | | | |
| Avoid abbreviations | ☐ | ☐ | |
| E.g., identifies a class as "Human Physiology" rather than "BIO 270" | | | |
| Yes - Free of abbreviations | | | |
| No - Abbreviation used but targeted reader will not recognize meaning | | | |
| N/A - Abbreviation used but targeted reader will recognize meaning | | | |

| | Yes | No | N/A |
|---|---|---|---|
| Measure between .5" and 1"<br>N/A - Not between .5 and 1" as a different margin width is recommended for a specific position, field, or major. | ☐ | ☐ | ☐ |

# References

Adams, R. L., & Morin, L. (1999). *Complete Résumé and Job Search Book for College Students*. Adams Media.

Bachman, L. F. (2005). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice, 21*(3), 5-18.

Black, P. (1998). Formative assessment: raising standards inside the classroom. *School Science Review, 80*, 39-46.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: Handbook I: Cognitive domain. *New York: David McKay, 19*, 56.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review; Psychological Review, 111*(4), 1061.

Bortoli, N. (1997). Résumés in the right: New rules make writing a winner easy. *Manage*, 49(1), 20-21

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*, 1-18.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34.

Brennan, R. L. (1997). A Perspective on the History of Generalizability Theory. *Educational Measurement: Issues and Practice, 16*(4), 14-20.

Bresciani, M. J. (2009). Assessment and evaluation. *Student Services: A Handbook for the Profession*, 4.

Brown, M. R., & Hayes, C. (1998). Résumé slip-ups. *Black Enterprise*, 28(1), 81-82.

Carver, D. S., & Smart, D. W. (1985). The effects of a career and self-exploration course for undecided freshmen. *Journal of College Student Personnel*.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.) (2008). Building a validity argument for the Test of English as a Foreign Language. London: Routledge

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an Argument-Based Approach to Validity Make a Difference?. *Educational Measurement: Issues and Practice*, *29*(1), 3-13.

Charney, D. H., Rayman, J., & Ferreira-Buckley, L. (1992). How Writing Quality Influences Readers' Judgments of Résumés in Business and Engineering. *Journal of Business and Technical Communication*, *6*(1), 38-74.

Cobb, G. W. (1998). The objective-format question in statistics: Dead horse, old bath water, or overlooked baby. In *annual meeting of the American Educational Research Association, San Diego, CA.*

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Lawrence Erlbaum.

Cronbach, L. J., Gleser, G. C., & Nanda, H. Rajaratnam. N.(1972). *The dependability of behaviour measurement: Theory of Generalizability for scores and profiles*.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391-418.

Crosby, O. (2009). *Résumés, Applications, and Cover Letters (2009)*. Bureau of Labor Statistics.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods, 3*, 412-423.

Dey, F., & Real, M. Adaptation of Casella's Model: Emerging Trends i C S i in Career Services.

Du, Y., & Wright, B. D. (1997). Effects of student characteristics in a large-scale direct writing assessment. *Objective measurement: Theory into practice,4*, 1-24.

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*.

Educational Testing Service (May, 2013). Glossary of standardized testing terms. Retrieved from http://www.ets.org/understanding_testing/glossary/

Embretson, S. E., & Hershberger, S. L. (Eds.). (1999). *The new rules of measurement: What every psychologist and educator should know*. Lawrence Erlbaum.

Engelhard Jr, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Engelland, B. T., Workman, L., & Singh, M. (2000). Ensuring service quality for campus career services centers: a modified SERVQUAL scale. *Journal of marketing Education*, *22*(3), 236-245.

Erwin, T. D. (1991). *Assessing student learning and development: A practical guide for college faculty and administrators*. San Francisco: Jossey-Bass Publishers.

Ewell, P. T. (2002). An emerging scholarship: A brief history of assessment. In T.W. Banta & Associates (Eds.), *Building a scholarship of assessment*. San Francisco: Jossey Bass Publishers.

Fallows, S., & Steven, C. (2000). Building employability skills into the higher education curriculum: a university-wide initiative. *Education+ training*, *42*(2), 75-83.

Farrokhi, F., & Esfandiari, R. (2011). A Many-facet Rasch Model to Detect Halo Effect in Three Types of Raters. *Theory and Practice in Language Studies*, *1*(11), 1531-1540.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp 237-270). Washington, DC: American Council on Education.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational researcher*, *18*(9), 27-32.

George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update. Allyn & Bacon. *Boston, USA*.

Girrell, K. W. (1979). The Poetry of Résumé Writing. *Journal of College Placement*, *39*(2), 49-50.

Groeber, J. F. (2006). *Designing and using rubrics for reading and language arts, K-6*. Corwin Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin 112*(3), 527.

Harcourt, J., & Krizan, A. C. (1989). A comparison of résumé content preferences of Fortune 500 personnel administrators and business communication instructors. *Journal of Business Communication*, *26*(2), 177-190.

Harvey, L., & Knight, P. T. (1996). *Transforming Higher Education*. Bristol, PA: Open University Press.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice, 10*(2), 33-41.

Hoheb, M. (2002). Résumé writing. *Scholastic Choices, 18*(3), 19-23.

Hornsby, J. S., & Smith, B. N. (1995). Résumé content: What should be included and excluded. *SAM Advanced Management Journal, 60*, 4-4.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. Journal of Research on Educational Effectiveness, 2(1), 88-110.

Hutchinson, K. L. (1984). Personnel administrators' preferences for résumé content: A survey and review of empirically based conclusions. *Journal of Business Communication, 21*(4), 5-14.

Hutchinson, K. L., & Brefka, D. S. (1997). Personnel administrators' preferences for résumé content: ten years after. *Business Communication Quarterly, 60*(2), 67-75.

Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology, 64*(5), 502.

Jennings, D. F., & Lumpkin, J. R. (1989). Functioning modeling corporate entrepreneurship: An empirical integrative analysis. *Journal of Management, 15*(3), 485-502.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin; Psychological Bulletin, 112*(3), 527.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance.*Educational measurement: issues and practice, 18*(2), 5-17.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3), 135-170.

Keeling, R. P., & Dungy, G. J. (2004). *Learning reconsidered: A campus-wide focus on the student experience*. National Association Of Student Personnel Administrators, American College Personnel Association.

Keeling, R. P. (2006). *Learning reconsidered 2: Implementing a campus-wide focus on the student experience*. American College Personnel Association.

Keeling, R. P., Wall, A. F., Underhile, R., & Dungy, G. J. (2008). Assessment reconsidered. *Washington, DC: National Association of Student Personnel Administrators*.

Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. *Personnel Psychology*, *34*(2), 263-289.

Klemp, G.O. (1977). Three factors of success in the world of work: implications for curriculum in higher education, paper presented to the 32nd *National Conference on Higher Education of the American Association for Higher Education.*

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-37.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice, 41*(4), 212-218.

Laker, D. R., & Laker, R. (2007). The five-year résumé: A career planning exercise. *Journal of Management Education*, *31*(1), 128-141.

Lamprianou, I. (2008). High stakes tests with self-selected essay questions: Addressing issues of fairness. *International Journal of Testing*, *8*(1), 55-89.

Lent, R. W., Larkin, K. C., & Hasegawa, C. S. (1986). Effects of a "Focused Interest" career course approach for college students. *Vocational Guidance Quarterly*, *34*(3), 151-159.

Linacre, J. M. (1989). Many-facet Rasch Measurement. Chicago: Measurement. *Evaluation, Statistics, and Assessment Press*.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, *3*(4), 486-512.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores.

Lovelace, H. W. (2001, May 7). Would you interview you? *Informationweek.com*, p. 163.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied measurement in education*, *3*(4), 331-345.

McDowell, E. E. (1987). Perceptions of the ideal cover letter and ideal résumé.*Journal of Technical Writing and Communication*, *17*(2), 179-191.

McNamara, T. (2000). *Language testing*. OUP Oxford.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice*, *11*(1), 3-9.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*(3), 293.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741.

Miner, T. (2000). Résumés as a Proactive Career Development Tool: An Innovation at Keuka College Career Center. *Career Planning and Adult Development Journal*, *16*(1), 72-79.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(10), 1-11.

Muraki, E., Hombo, C. M., & Lee, Y. W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, *24*(4), 325-337.

Murphy, K. R., & Cleveland, J. N. (1991). *Performance appraisal: An organizational perspective*. Allyn & Bacon.

Myford, C. M., & Wolfe, E. W. (2002). When raters disagree, then what: examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement*, *3*(3), 300-324.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4, 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5, 189–227.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371-389.

National Association of Student Personnel Administrators & American College Personnel Association (2004). Learning reconsidered: A campus-wide focus on the student experience. Washington, D.C.

Nichols, J. (2001). Building the perfect résumé. *Careers & Colleges*, 21(3), 41-43

Nichols, A. (2007). Causal inference with observational data. *Stata Journal*,*7*(4), 507.

Nitko, A. J. (1996). *Educational assessment of students*. Prentice-Hall Order Processing Center, PO Box 11071, Des Moines, IA 50336-1071.

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merrill.

New York State Department of Education Office of State Assessment (2013). New York State Education Department test development process. Retrived from http://www.p12.nysed.gov/assessment/teacher/

Orem, C. D. (2012). Demonstrating validity evidence of meta-assessment scores using generalizability theory (Doctoral dissertation, James Madison University).

Osterlind, S. J. (1997). *Constructing test items: Multiple-choice, constructed-response, performance and other formats*. Springer.

Pascarella, E. T., & Terenzini, P. T. (2005). *How College Affects Students: A Third Decade of Research. Volume 2*. Jossey-Bass, An Imprint of Wiley.

Palomba, C. A., & Banta, T. W. (1999). *Assessment Essentials: Planning, Implementing, and Improving Assessment in Higher Education. Higher and Adult Education Series*. Jossey-Bass, Inc. Publishers.

Pearl, J. (2000). *Causality: models, reasoning and inference* (Vol. 29). MIT Press.

Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of marketing research*, 6-17.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 221-262). New York: Macmillan.

Petrecca, L. (2010, April). Grads' toughest test? Job market. *USA Today*. B1-B2.

Pibal, D. C. (1985). Criteria for effective résumés as perceived by personnel directors. *Personnel Administrator*, *30*(5), 119-123.

Popham, W. J. (1997). What's wrong-and what's right-with rubrics. *Educational Leadership*, *55*, 72-75.

Rampell, C. & Hernandez, J. C. (2010, April 2). Signaling jobs recovery, payrolls surged in March. *New York Times*.

Raymond, M. R., & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement*, *30*(3), 253-268.

Raymond, M. R., Harik, P., & Clauser, B. E. (2011). The Impact of Statistically Adjusting for Rater Effects on Conditional Standard Errors of Performance Ratings. *Applied Psychological Measurement*, *35*(3), 235-246.

Ross, C. M., & Young, S. J. (2005). Résumé Preferences Is It Really "Business as Usual"?. *Journal of Career Development*, *32*(2), 153-164.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*,*88*(2), 413.

Savickas, M. L. (2010). Vocational counseling. *Corsini Encyclopedia of Psychology*.

Schramm, R. M., & Dortch, R. N. (1991). An Analysis of Effective Résumé Content, Format, and Appearance Based on College Recruiter Perceptions.*Bulletin of the Association for Business Communication*, *54*(3), 18-23.

Schuh, J. H., & Upcraft, M. L. (2001). *Assessment Practice in Student Affairs: An Application Manual. The Jossey-Bass Higher and Adult Education Series*. Jossey Bass Publishers, 350 Sansome St., San Francisco, CA 94104.

Schuh, J. H. (2009). *Assessment methods for student affairs*. Jossey-Bass.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental design for generalized causal inference. Boston: Houghton-Mifflin.

Shakoor, A. T. (2001). Developing a professional résumé and cover letter that work. *Black Collegian*, 32, 16.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215-232.

Shavelson, R. J. (2009). *Measuring college learning responsibly: Accountability in a new era*. Stanford University Press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications, Incorporated.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. American Psychologist, 44(6), 922-932.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational researcher*, 4-14.

Shutt, M. D., Garrett, J. M., Lynch, J. W., & Dean, L. A. (2012). An Assessment Model as Best Practice in Student Affairs. *Journal of Student Affairs Research and Practice, 49*(1), 65-82.

Simon, A. (2010, May). Job outlook for college grads better than last year. *The Greenville News*, A1, A4.

Smith, R. M., Julian, E., Lunz, M., Stahl, J., Schulz, M., & Wright, B. D. (1994). Applications of conjoint measurement in admission and professional certification programs. *International Journal of Educational Research, 21*(6), 653-664

Stanley-Weigand, P. (1991). Organizing the Writing of Your Résumé. *Bulletin of the Association for Business Communication*, *54*(3), 11-12.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*(3), 239-261.

Sunshine, R. A. (2010, February). *The Budget and Economic Outlook: Fiscal Years 2010 to 2020*. Congressional Budget Office. Retrieved from http://cbo.gov/ftpdocs/108xx/doc10871/Chapter2.shtml#1105094

Suskie, L. (2006). Accountability and quality improvement. *Revisiting outcomes assessment in higher education. Westport, CT: Libraries Unlimited.*

Suskie, L. (2009). *Assessing student learning: A common sense guide*. Jossey-Bass.

Thomas, J. H., & McDaniel, C. R. (2004). Effectiveness of a required course in career planning for psychology majors. *Teaching of Psychology, 31*(1), 22-27.

Tillotson, K., & Osborn, D. (2012). Effect of a résumé-writing workshop on résumé-writing skills. *Journal of Employment Counseling*, *49*(3), 110-117.

Toporek, R. L., & Flamer, C. (2009). the résumé's secret identity: a tool for narrative exploration in multicultural career counseling. *Journal of Employment Counseling*, *46*(1), 4-17.

Tsai, W. C., Chi, N. W., Huang, T. C., & Hsu, A. J. (2011). The effects of applicant résumé contents on recruiters' hiring recommendations: The mediating roles of recruiter fit perceptions. *Applied Psychology*, *60*(2), 231-254.

UNESCO (1998). World education report: Teachers and teaching in a changing world. Paris: UNESCO.

Upcraft, M. L., & Schuh, J. H. (1996). *Assessment in Student Affairs: A Guide for Practitioners. The Jossey-Bass Higher and Adult Education Series*. Jossey-Bass Inc., Publishers, 350 Sansome St., San Francisco, CA 94104.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. *Handbook of statistics*, *26*, 81-124.

Wiggins, G. (1989) "A True Test: Toward More Authentic and Equitable Assessment," Phi Delta Kappan, 70, 9 (May).

Wiggins, G. P. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. Jossey-Bass.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*, 35-51.

Wolfe, E. W., & Dobria, L. (2008). Applications of the multifaceted Rasch model. *Best practices in quantitative methods*, 71-85.

Wolfe, E. W. (2009). Item and rater analysis of constructed response items via the multi-faceted Rasch model. *Journal of Applied Measurement 10*(3), 335-347.

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College composition and communication, 50*(3), 483-503.